Statistical Analysis Plan	
Study Code	D933AC00001
Edition Number	6.0
Date	12 October 2021

A Phase III Randomized, Double-Blind, Placebo-Controlled, Multi-Regional, International Study of Durvalumab in Combination with Gemcitabine plus Cisplatin versus Placebo in Combination with Gemcitabine plus Cisplatin for Patients with First-Line Advanced Biliary Tract Cancers (TOPAZ-1)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

Abbreviation or special term	Explanation
ADA	Anti-drug antibody
AE	Adverse event
AEPI	Adverse event of possible interest
AESI	Adverse event of special interest
ALP	Alkaline phosphatase
ALT	Alanine aminotransferase
AoV	Ampulla of Vater
AST	Aspartate aminotransferase
ATC	Anatomical therapeutic chemical
AUC	Area under the curve
AZ	AstraZeneca
BICR	Blinded independent central review
BLQ	Below limit of quantification
BoR	Best objective response
BP	Blood pressure
BTC	Biliary tract cancer
CI	Confidence interval
CMH	Cochran-Mantel Haenszel
CR	Complete response
CrCl	Creatine clearance
CRO	Contract research organization
CRF	Case report form
CSP	Clinical study protocol
CSR	Clinical study report
CT	Computed tomography
CTCAE	Common terminology criteria for adverse event
ctDNA	Circulating tumor deoxyribonucleic acid
CTLA-4	Cytotoxic T-lymphocyte-associated antigen-4
CV	Coefficient of variation
DAE	Discontinuation of investigational produce due to adverse event
DBL	Database lock
DCO	Data cut-off

Abbreviation or special term	Explanation
DCR	Disease control rate
DCR-24w	DCR at 24 weeks
DCR-32w	DCR at 32 weeks
DCR-48w	DCR at 48 weeks
DoR	Duration of response
d.p.	Decimal place
ECG	Electrocardiogram
ECOG	Eastern Cooperative Oncology Group
eCRF	Electronic case report form
EORTC	European Organization for Research and Treatment of Cancer
EQ-5D	EuroQol 5-dimension
EQ-5D-5L	EuroQol 5-dimension, 5 level health state utility index
EQ-VAS	EuroQol visual analogue scale
FA	Final analysis
FAS	Full analysis set
FAS-32w	Subjects from Full analysis set with an opportunity for at least 32 weeks of follow-up at the time of IA-1 DCO
FH	Fleming-Harrington
FWER	Familywise error rate
GB	Gallbladder
Gem/Cis	Gemcitabine plus cisplatin
HL	Hy’s Law
HLGT	High level group term
HLT	High level term
HOSPAD	Hospital resource use module
HR	Hazard Ratio
HRQoL	Health-related quality of life
IA-1/2	Interim analysis 1/2
ICU	Intensive care unit
IDMC	Independent data monitoring committee
IHC	Immunohistochemistry
imAE	Immune-mediated adverse event
IP	Investigational produ
IPD	Important protocol deviation

Abbreviation or special term	**Explanation**
IRC	Independent review charter
ITT	Intention to treat
IV	Intravenous
IVRS	Interactive voice response system
IWRS	Interactive web response system
KM	Kaplan-Meier
LD	Longest diameter
LLOQ	Lower limit of quantification
MedDRA	Medical dictionary for regulatory activities
MMRM	Mixed model repeated measures
MRI	Magnetic resonance imaging
MSI	Microsatellite instability
NA	Not applicable
nAB	Neutralizing antibody
NC	Not calculable
NCI	National Cancer Institute
NE	Not evaluable
NED	No evidence of disease
NQ	Not quantifiable
NR	Not reported
NS	No sample
NTL	Non-target lesions
OAE	Other significant adverse events
OS	Overall survival
ORR	Objective response rate
PAP	Payer Analysis Plan
PD	Progressive disease
PD-1	Programmed cell death 1 (CD279)
PD-L1	Programmed cell death ligand-1 (also known as B7 homolog 1, CD274)
PFS	Progression free survival
PGIS	Patient global impression of severity
PK	Pharmacokinetics
PR	Partial response
PRO	Patient reported outcome

Abbreviation or special term	**Explanation**
PRO-CTCAE	Patient Reported Outcomes Common Terminology Criteria for Adverse Events
PS	Performance status
PT	Preferred term
q3w	Every 3 weeks
q4w	Every 4 weeks
QLQ-C30	30-Item Core Quality of Life Questionnaire
QLQ-BIL21	21-Item Cholangiocarcinoma and Gallbladder Cancer Quality of Life Questionnaire
QoL	Quality of Life
QTcF	QT interval corrected for heart rate using Friderica's formula
qXw	Every X weeks
RDI	Relative dose intensity
RECIST	Response evaluation criteria in solid tumors
REML	Restricted maximum likelihood
SAE	Serious adverse event
SAF	Safety analysis set
SAP	Statistical analysis plan
SD	Stable disease
SoC	Standard of care
SMQ	Standardized MedDRA Queries
T3	Triiodothyronine
T4	Thyroxine
TEAE	Treatment emergent adverse event
TIP	Tumor and/or immune cell positivity
TL	Target lesion
TMB	Tumor mutation burden
TSH	Thyroid stimulating hormone
TTD	Time to deterioration
ULN	Upper limit of normal
VAS	Visual analogue scale
WHO	World Health Organisation
WHO-DD	World Health Organisation drug dictionary

AMENDMENT HISTORY

Category: Change refers to	Date	Description of change	In line with CSP? Y (version) / N / NA	Rationale
Primary or secondary endpoints	14 Aug 2020	Text update for analysis of OS and PFS (Table 16, Sections 4.2.2 and 4.2.3.1).	NA	To improve previous intent
	14 Aug 2020	Text updated for DoR and DCR to include additional analyses (Table 16, Sections 3.2.7, 4.2.3.3 and 4.2.3.4).	NA	To improve previous intent and align with TFLs
	14 Aug 2020	Text updated for PD-L1 and MSI subgroups (Section 4.2.2).	Y (v4.0)	To align with CSP
	14 Jan 2021	IA-1 outcomes for ORR, BoR and DoR added to Table 5.	NA	To improve previous intent
	14 Jan 2021	Clarification of outcomes required for IA-1 (Sections 3.2.3, 3.2.4, 3.2.5, 3.2.6, 4.2.3.2, 4.2.3.3 and 5.1.1).	NA	To improve previous intent
	14 Jan 2021	Clarification that ORR, BoR and DoR will be summarized both by unconfirmed and	NA	To improve previous intent and align with TFLs

		confirmed responses and modification of text to clarify analysis (Sections 3.2.3, 3.2.4, 3.2.5, 4.2.3.2, 4.2.3.3 and 5.1.1).		
	14 Jan 2021	Minimum efficacy criteria for IA-1 updated and URC criteria for submission recommendation added (Section 5.1.1.).	NA	To improve previous intent
	30 Jun 2020	Clarification that ORR, DoR, BoR, DCR and tumor size analyzed for BICR assessments at IA-1 only (Sections 3.2.3, 3.2.4, 3.2.5, 3.2.6 and 3.27).	NA	To improve previous intent
Sample size and multiple testing procedure	14 Aug 2020	Sample size updated. (Section 1.3).	Y (v4.0)	To align with CSP
	14 Aug 2020	Clarification of the number of randomized subjects from China (Section 1.3).	Y (v4.0)	To align with CSP
	14 Aug 2020	Update to multiple - testing procedure and timing of IA-1 and IA-2 (Sections 1.3, 3.1.4, 3.2.3, 3.2.4, 4, 4.2.1,	Y (v4.0)	To align with CSP and TFLs

		4.2.3.3, 5.1.1, 5.1.2, Table 17 and Figure 2).		
	14 Jan 2021	Update to multiple testing procedure and timing of IA-2 (Sections 1.3, 4, 4.2.1, 5.1.2 and Figure 2).	Y (v6.0)	To align with CSP
	5 Mar 2021	Section 1.3: Included sample size calculations for FH(0, 1)	Y (v7.0)	To align with CSP
	5 Mar 2021	Section 4.2.1: describe calculation of significance level at FA for FH(0, 1) test. Removed MTP figure.	Y (v7.0)	To align with CSP
	11 Oct 2021	Text updated on significance level determinations at IA2 and FA (Sections 1.3, 4.2.1 and 5.1.2).	NA	To improve previous intent
Derivation of primary or secondary endpoints	14 Aug 2020	Step 3 of the TL visit response subsequent to CR updated (Section 3.1.1).	NA	Clarification of derivation
	14 Aug 2020	Text updated to include derivation of OS in absence of survival call (Section 3.2.1).	NA	To improve previous intent
	14 Aug 2020	Text modified on recording of new	NA	To improve previous intent

	lesions and therefore also overall visit response. (Section 3.1.3 and Table 8).		
14 Aug 2020	PFS 2 missed visit clarification, that NE visit is not considered as missed visit (Section 3.2.2).	NA	To improve previous intent
14 Aug 2020	Clarification of derivation of best percentage change in tumor size (Section 3.2.6).	NA	Clarification of derivation
14 Aug 2020	Text updated for DCR (Abbreviations and Section 3.2.7)	NA	To improve previous intent
14 Aug 2020	Clarification of visit assessments to be included for PRO summaries and analysis (Section 3.3).	NA	To improve previous intent
14 Aug 2020	Details on scoring of EORTC QLQ-C30 added (Section 3.3.1 and Table 9).	NA	To improve previous intent
14 Aug 2020	Addition of table describing two missed visits for EORTC QLQ-C30/BIL21 (Section 3.3.1 and Table 12).	NA	To improve previous intent

14 Aug 2020	Clarification of event and censoring rules for time to deterioration of PRO outcomes (Table 13).	NA	To improve previous intent
14 Aug 2020	Removal of overall compliance and clarification of definitions for expected and evaluable questionnaires for PRO outcomes (Section 3.3.6).	NA	To improve previous intent
14 Aug 2020	Text on visit windows and visit updated (Section 3.5.12.2).	NA	To improve previous intent
14 Aug 2020	Text added to clarify that visit data will only be summarized if number of observations>=20 in at least one treatment arm (Section 3.5.12.2).	NA	To improve previous intent
14 Aug 2020	Text modified for rules for handling missing data (Section 3.5.12.3).	NA	To improve previous intent
14 Aug 2020	MMRM model for PRO outcomes updated to include baseline score by visit interaction (Section 4.2.4.1).	NA	To improve previous intent

14 Aug 2020	Clarification of ORR (Section 4.2.3.2).	Y (v3.0)	To align with CSP.
14 Jan 2021	Clarification of two missed visit rules for RECIST assessments (Section 3.2.2).	NA	To improve previous intent
14 Jan 2021	Additional DCR endpoint added for 48 weeks (Abbreviations, Sections 3.27 and 4.2.3.4, Table 17)	NA	To align with TFLs
14 Jan 2021	Clarification of event and censoring rules for time to deterioration of PRO outcomes (Table 13).	NA	To improve previous intent
14 Jan 2021	Definition of subset of population for time to symptom deterioration added for EORTC QLQ-C30 (Section 3.3.1)	NA	To improve previous intent
14 Jan 2021	Symptom improvement rate definition added for EORTC QLQ-BIL21 (Section 3.3.2)	NA	To improve previous intent
14 Jan 2021	Minor text updates to compliance section (Section 3.3.6).	NA	To improve previous intent
14 Jan 2021	Clarification that exact 95% CI for ORR will be computed using	NA	To improve previous intent

	Clopper-Pearson for IA-1 (Section 5.1.1).		
5 Mar 2021	Clarified study day calculation (Section 4.1)	NA	To improve previous intent
5 Mar 2021	Added Cockroft-Gault formula for CrCl calculation (Section 3.5.7)	NA	To improve previous intent
5 Mar 2021	Section 3.2.3 – Updated text around ORR analyses for IA-1	NA	To improve previous intent
5 Mar 2021	Section 4.2.6.1: Clarified definition of first subsequent anticancer therapy date. Added stenting events analysis	NA	To improve previous intent
5 Mar 2021	Section 3.2.7 clarified denominator for DCR analysis, and the subset for IA-1 analysis.	NA	To improve previous intent
5 Mar 2021	Updated Table 9 to cover all possible cases of RECIST response in ITT	NA	To improve previous intent
5 Mar 2021	Section 4.2.9 – defined cut off for PD-L1 analyses.	NA	To improve previous intent
5 Mar 2021	Added Section 4.2.3.4 for BoR as it was missing.	NA	To improve previous intent

	07 Jun 2021	Text added for Time to symptom deterioration (Section 3.3.1).	NA	To improve previous intent
	07 Jun 2021	Addition of rows to table describing two missed visits for EORTC QLQ-C30/BIL21 for follow-up visits (Section 3.3.1 and Table 13).	NA	To improve previous intent
	07 June 2021	Clarification of order of responses for BoR (Section 3.2.5).	NA	To improve previous intent
	07 June 2021	Updated text to use randomization date as reference for DCR (Section 3.2.6).	NA	To improve previous intent
	07 Jun 2021	Amended baseline PRO to be last prior to first dose (Section 4.1)	Y (v7.0)	To align with CSP
Data presentation	14 Aug 2020	Text modified for list of safety and exploratory endpoints (Tables 3 and 4).	Y (v3.0/v4.0)	To align with CSP
	14 Aug 2020	Updated vital signs text to clarify multiple timepoints per visit (Section 3.5.8).	NA	To improve previous intent
	14 Aug 2020	Clarification of OS analysis to include OS	NA	To align with TFLs

	at 12 and 18 months (Section 4.2.2).		
14 Aug 2020	Additional text added to include Max-Combo analysis for OS/PFS (Section 4.22).	NA	To improve previous intent
14 Aug 2020	Clarification of PFS analysis to include PFS at 6, 9 and 12 months (Section 4.2.3.1).	NA	To align with TFLs
14 Aug 2020	List of summaries for percentage change from baseline in tumor size added (Section 4.2.3.5).	NA	To align with TFLs
14 Aug 2020	Listing of all RECIST 1.1 data and summary of new lesions added (Section 4.2.3.6).	NA	To align with TFLs
14 Aug 2020	Symptoms updated for data presentations of EORTC QLQ-C30 (Section 4.2.4.1).	NA	To align with TFLs
14 Aug 2020	Symptoms updated for data presentations of EORTC QLQ-BIL21 (Section 4.2.4.2).	NA	To align with TFLs
14 Aug 2020	Updated list of AE summaries (Section 4.2.6.1).	NA	To align with TFLs
14 Aug 2020	AE section updated to include summaries for AESI/AEPI and	NA	To improve previous intent

	Infection AEs separately by HLGT/HLT and custom pooled terms (Abbreviations and Section 4.2.6.2).		
14 Aug 2020	Text on deaths modified (Section 4.2.6.1).	NA	To align with TFL
14 Jan 2021	Addition of descriptive statistics for tumor size by visits and clarification that summaries will be produced for both Investigator and BICR assessments (Section 4.2.3.5).	NA	To improve previous intent
14 Jan 2021	Summaries of RECIST assessments added and listing for all RECIST data removed (Section 4.2.3.6).	NA	To align with TFLs
14 Jan 2021	Addition of summary comparing BoR between Investigator and BICR assessments (Sections 4.2.3.6 and 5.1.1).	NA	To align with TFLs
14 Jan 2021	Summaries for OAE and for Infusion reaction adverse events added (Sections 4.2.6.1 and 4.2.6.2).	NA	To improve previous intent and to align with TFLs

14 Jan 2021	Section on infection AEs removed as no longer required to be reported separately (Section 4.2.6.2).	NA	To improve previous intent
14 Jan 2021	Abnormal Thyroid function summaries updated to include free T3 and free T4 (Abbreviations and section 4.2.6.4).	NA	To improve previous intent
14 Jan 2021	New section for COVID-19 added (Section 4.2.12).	NA	To document potential effects of COVID-19 on trial
5 Mar 2021	Added a listing for MSI and PD-L1 data (Section 4.2.6.4)	NA	To improve previous intent
5 Mar 2021	Added back paragraphs about Thyroid function summaries (removed by mistake) (Section 4.2.6.4)	NA	To improve previous intent
5 Mar 2021	Included CrCl summaries using Cockroft-Gault formula (Section 4.2.6.4)	NA	To improve previous intent
5 Mar 2021	Section 4.2.3.2 – ORR analysed using CMH test, and in FAS population	N	To follow FDA feedback of 25th Jan 2021

5 Mar 2021	Section 4.2.3.6: BICR vs Investigator analysis will be produced for IA-1 both in FAS and FAS with measurable disease at baseline	NA	To improve previous intent
5 Mar 2021	Section 4.2.6.1 – Other significant adverse events analysis repeated for AEs related to study treatment	NA	To improve previous intent
5 Mar 2021	Section 3.5.6 clarified that AESI will be classified based on PTs, not higher level terms.	NA	To improve previous intent
5 Mar 2021	Section 4.2.10: specified that medical and surgical history will be presented, to reflect actual CSR analyses.	NA	To improve previous intent
5 Mar 2021	Section 5.1.1: added CMH test for ORR, and listed sensitivity analyses in FAS-32w subset	N	To follow FDA feedback of 25th Jan 2021
5 Mar 2021	Section 3.6: updated based on latest SOP	NA	To improve previous intent
5 Mar 2021	Section 3.2.3: mentioned that ORR will be analysed in	NA	To improve previous intent

	FAS as a sensitivity analysis		
5 Mar 2021	Section 4.2, Section 4.2.2. Changed OS testing method at FA to FH(0,1) and including log-rank test as a sensitivity at FA. Added rationale for FH(0, 1) test at FA	Y (V7.0)	To reflect changes in CSP
30 Jun 2021	Clarification that time to response and number of subjects still in response will also be summarized for DoR (Section 4.2.3.3).	NA	To improve previous intent
07 Jun 2021	Pregnancy listing added (Section 4.2.6.4).	NA	To improve previous intent and align with TFLs
07 Jun 2021	Clarification that HR for OS at FA will be calculated from Cox proportional hazard model (Table 18).	Y (v7.0)	To reflect changes to CSP
07 Jun 2021	Text added for adjusted alpha for OS and PFS (Section 4.2.2).	NA	To improve previous intent
07 Jun 2021	Clarification that ECOG is reported at screening (Section 4.2.2 and 4.2.10).	NA	To improve previous intent

	07 Jun 2021	eCDF/PDF sections removed from PRO analysis (Sections 4.2.4.1, 4.2.4.2 and 4.2.4.4).	NA	To improve previous intent and align with TFLs
	07 Jun 2021	Text modified for TTD for PROs to include Cox proportional hazard model for HR (Section 4.2.4.1).	NA	To improve previous intent
	07 Jun 2021	Text amended to clarify that for IA-1 DoR only analyzed in measurable disease subset (Section 5.1.1)	NA	To improve previous intent
Other	14 Aug 2020	For consistency replaced all occurrences of “patient(s)” in body of text with “subject(s)”. Applies throughout	NA	To improve previous intent
		Updated CSP version number from 2.0 to 4.0 (Section 1).	Y (v4.0)	To align with CSP
	14 Aug 2020	Minor updates to study design text (Section 1.2).	Y (v3.0)	To align with CSP
	14 Aug 2020	Update to study design diagram (Figure 1).	Y (v3.0)	To align with CSP
	14 Aug 2020	Clarification of text for health care resource use variables and	NA	To improve previous intent

	analysis (Sections 3.4 and 4.2.5).		
14 Aug 2020	Text corrected to state deaths should be reported for FAS (Section 3.5).	NA	To improve previous intent
14 Aug 2020	Study treatments table updated (Table 14).	Y (v3.0/v4.0)	To align with CSP
14 Aug 2020	Derivation of actual exposure for durvalumab or placebo corrected to include dose delays instead of dose interruptions (Section 3.5.2.1).	NA	Correction of derivation
14 Aug 2020	Text updated to clarify dose interruptions and dose delays for durvalumab or placebo (Section 3.5.2.1).	NA	To improve previous intent
14 Aug 2020	Text updated to remove protocol window of 3 days from calculation on duration of dose delays for durvalumab /placebo (Section 3.5.2.1).	NA	To improve previous intent
14 Aug 2020	Calculation for duration of gemcitabine/ cisplatin dose delays corrected (Section 3.5.2.2).	NA	Correction to derivation

		Text updated to clarify dose reductions, interruptions and dose delays for gemcitabine or cisplatin (Section 3.5.2.2).	NA	To improve previous intent
	14 Aug 2020	Text updated to clarify calculation of RDI (Section 3.5.3)	NA	To improve previous intent
	14 Aug 2020	AE text modified to clarify on treatment AEs exclude AEs after subsequent anti-cancer therapy (Section 3.5.4).	NA	To improve previous intent
	14 Aug 2020	Derivation of RR added (Section 3.5.10).	NA	To improve previous intent
	14 Aug 2020	Text modified on AESI to include AEPI. (Section 3.5.6 and 4.2.6.2).	NA	To improve previous intent
	14 Aug 2020	Text modified text on reference ranges used for laboratory variables to indicate project ranges will be used (Section 3.5.7).	NA	To improve previous intent
	14 Aug 2020	Text corrected to state healthcare resource use should be reported for safety analysis set (Section 4.2.5).	NA	To improve previous intent

14 Aug 2020	Text modified for infection AEs to included HLGT/HLT pooled terms and custom pooled terms (Section 4.2.6.2).	NA	To improve previous intent
14 Aug 2020	Definition of Hy's law updated (Section 4.2.6.4).	Y (v3.0)	To align with CSP
14 Aug 2020	Text modified on PK data (Section 4.2.7).	Y (v3.0)	To align with CSP
14 Aug 2020	Stratification factors added to list of baseline summaries (Section 4.2.10).	NA	To improve previous intent
14 Aug 2020	Text modified on concomitant medication and other treatment summaries (Section 4.2.11).	NA	To improve previous intent
14 Aug 2020	References updated (Section 7).	NA	To improve previous intent
14 Jan 2021	Updated CSP version number from 4.0 to 6.0 (Section 1).	Y (v6.0)	To align with CSP
14 Jan 2021	Definition of FAS for IA-1 including only subjects with at least 32 weeks follow-up added (Section 2.1.1).	NA	To improve previous intent
14 Jan 2021	Data rules added for data below limit of	NA	To improve previous intent

	quantification for PK analysis (Section 3.6).		
14 Jan 2021	Removal of redundant text (Sections 3.9.1 and 4.2.2).	NA	To improve previous intent
5 Mar 2021	Corrected spelling in CMH	NA	To improve previous intent
5 Mar 2021	Section 4.2.3.3 – corrected Section reference	NA	To improve previous intent
07 Jun 2021	Definition of PRO analysis set added (Section 2.1.2).	NA	To improve previous intent
07 Jun 2021	Appendix B added - definition of visit windows and referenced in Sections 3.2.7 and 4.2.6.1.	NA	To improve previous intent
30 Jun 2021	Update of HRQoL to global health status/QoL throughout and removal of detailed description of items included for EORTC QLQ-C30 global health status/QoL (Sections 3.31 and 4.2.4.2 and Table 18).	NA	To improve previous intent
30 Jun 2021	Clarification that reporting of compliance data for PROs by visit should	NA	To improve previous intent

		not be restricted to n≥20 subjects in one of the treatment arm (Section 3.3.6).		
	30 Jun 2021	Text updated for Immune-mediated adverse events (Section 3.5.6).	NA	To improve previous intent
	07 Jun 2021	Changes to the analysis from CSP added to Section 6.	NA	To improve previous intent
	07 Jun 2021	Text for NR and NS added to BLQ rules for PK data (Section 3.6).	NA	To improve previous intent
	30 Jun 2021	Clarification of definition of Treatment emergent ADA positive (Section 3.7).	NA	To improve previous intent
	30 Jun 2021	Clarification that that study will have met its primary objective if Arm A is statistically significantly superior to Arm B, either at IA-2 or at the final analysis (Section 4).	Y (v7.0)	To align with CSP
	30 Jun 2021	Text added to state that subgroup analysis will not be performed for subgroups with < 5 subjects (Section 4.1).	NA	To improve previous intent
	30 Jun 2021	Definition of baseline moved to sub-section	NA	To improve previous intent

		of General principles Section 4.1.1.		
	30 Jun 2021	General considerations for safety and PRO assessments including visit window and missing data rules moved to new Section 4.2.6.1.	NA	To improve previous intent
	30 Jun 2021	Hematopoietic SMQs added to other significant events (Section 4.2.6.2).	NA	To improve previous intent
	30 Jun 2021	Additional text added to clarify PD-L1 expression (low and high) and MSI status (high and stable) (Section 4.2.9 and Table 19).	NA	To improve previous intent
	30 Jun 2021	Virology status added to list of baseline characteristics (Section 4.2.10).	NA	To improve previous intent

1 STUDY DETAILS

This statistical analysis plan (SAP) contains a more detailed description of the analyses in the clinical study protocol (CSP) and is based on version 7.0 of the CSP.

This SAP will apply to the phase III study to evaluate the clinical benefit of adding durvalumab to the established chemotherapy regimen of gemcitabine and cisplatin for the treatment of subjects with previously untreated, unresectable locally advanced or metastatic biliary tract cancer (BTC).

The target population includes subjects ≥18 years of age with previously untreated, unresectable locally advanced or metastatic BTC. Cancer of Ampulla of Vater (AoV) has a different genetic profile than other subtypes of BTC and therefore to minimize the diversity of the study population will be excluded from the study. Subjects must have at least 1 lesion that qualifies as a Response Evaluation Criteria in Solid Tumors (RECIST) 1.1 Target Lesion (TL) at baseline and also have a World Health Organization (WHO)/Eastern Cooperative Oncology Group (ECOG) performance status (PS) of 0 or 1 at enrolment.

Refer to the CSP for a detailed description of the rationale for this study as well as its inclusion/exclusion criteria.

1.1 Study objectives

1.1.1 Primary objectives

The primary study objective and the corresponding Endpoint/Variable for this study are shown in Table 1.

Table 1: Primary study objective and corresponding Endpoint/Variable

Objective	Endpoint/Variable
To assess the efficacy of Arm A compared to Arm B in terms of OS in patients with first-line advanced BTC	OS: Time from date of randomization until date of death by any cause

Note: Subjects in Arm A will receive durvalumab plus gemcitabine/cisplatin combination therapy; subjects in Arm B will receive placebo plus gemcitabine/cisplatin therapy. BTC Biliary tract cancer; OS Overall survival.

1.1.2 Secondary objectives

The secondary study objectives and the corresponding endpoints/variables for this study are shown in Table 2.

Table 2: Secondary study objectives and corresponding endpoints/variables

Objectives	Endpoint/Variables
To further assess the efficacy of Arm A compared to Arm B in terms of PFS, ORR, and DoR in patients with first-line advanced BTC	Endpoints based on investigator assessment according to RECIST 1.1: • PFS: Time from date of randomization until tumor progression or death due to any cause • ORR: The percentage of evaluable patients with investigator-assessed visit response of CR or PR • DoR: Time from first documented response (CR or PR) until date of documented progression or death in the absence of disease progression
For IA-1: To summarize the efficacy of Arm A compared to Arm B in terms of ORR and DoR in patients with first-line advanced BTC	ORR and DoR according to RECIST 1.1 using BICR assessments
To assess disease-related symptoms, impacts, and HRQoL in patients treated with Arm A compared to Arm B	EORTC QLQ-C30: Global health status/QoL and impacts (e.g., physical function); multi-term symptoms (e.g., fatigue); and single items (e.g., appetite loss, insomnia) EORTC QLQ-BIL21: Single-item symptoms (e.g., abdominal pain [item 42], pruritus [item 36], jaundice [item 35])
To assess the efficacy of Arm A compared to Arm B by PD-L1 expression	Association of PD-L1 expression level with PFS, ORR, DoR, and DCR according to RECIST 1.1 using Investigator assessments and OS
To assess the PK of durvalumab when used in combination with gemcitabine/cisplatin	Serum concentration of durvalumab (peak and trough concentrations)

To investigate the immunogenicity of durvalumab	Reporting tiered results of ADAs for durvalumab

Note: Subjects in Arm A will receive durvalumab plus gemcitabine/cisplatin combination therapy; patients in Arm B will receive placebo plus gemcitabine/cisplatin therapy.
ADA Anti-drug antibody; BICR Blinded independent central review; BTC Biliary tract cancer; DCR Disease control rate; DoR Duration of response; EORTC European Organization for Research and Treatment of Cancer; HRQoL Health-related quality of life; IA-1 Interim analysis-1; ORR Objective response rate; OS Overall survival; PD-L1 Programmed cell death ligand 1; PFS Progression free survival; PK Pharmacokinetics; QLQ-BIL21 21-Item Cholangiocarcinoma and Gallbladder Cancer Quality of Life Questionnaire QLQ-C30 30-Item Cancer Quality of Life Questionnaire; QoL Quality of Life Questionnaire; RECIST Response Evaluation Criteria in Solid Tumors.

1.1.3 Safety objective

The safety study objective and the corresponding Endpoint/Variable for this study is shown in Table 3.

Table 3: Safety study objective and corresponding Endpoint/Variable

Objective	Endpoint/Variable
To assess the safety and tolerability profile of Arm A compared to Arm B in patients with first-line advanced BTC	AEs, physical examinations, laboratory findings, WHO/ECOG PS, ECG and vital signs

Note: Subjects in Arm A will receive durvalumab plus gemcitabine/cisplatin combination therapy; subjects in Arm B will receive placebo plus gemcitabine/cisplatin therapy.
AE Adverse event; BTC Biliary tract cancer; ECOG Eastern Cooperative Oncology Group; PS Performance status; WHO World Health Organization.

1.1.4 Exploratory objectives

The exploratory study objectives and the corresponding endpoints/variables for this study are shown in Table 4.

Table 4: Exploratory study objective and corresponding endpoints/variables

Objectives	Endpoint/Variables
To investigate the efficacy of Arm A compared to Arm B by candidate biomarkers (for example but not limited to TMB and MSI) that may	Association of candidate biomarkers including, but not limited to TMB, MSI and/or tumor mutations with: OS and PFS, ORR, DoR and DCR according to RECIST 1.1 using Investigator assessments

correlate with drug activity or identify patients likely to respond to treatment	
To evaluate circulatory-based biomarkers and associations with efficacy parameters, including, but not limited to, ctDNA. It is not applicable in China	Association with circulatory-based biomarkers including, but not limited to, ctDNA-based TMB, whole blood gene expression, etc, with efficacy assessments
To assess patient reported treatment tolerability using PRO-CTCAE and global assessment of treatment tolerability	PRO-CTCAE (pre-selected items based on treatment arms) and global assessment of treatment tolerability (QLQ-BIL21 item 49)
To assess the patients' global impression of the severity of cancer symptoms	PGIS
To explore the impact of treatment and disease state on health state utility using the EQ-5D-5L	The EQ-5D-5L health state utility instrument will be used to derive health state utility based on patient reported data
To explore the impact of treatment and disease on healthcare resource use	The HOSPAD module will be used to collect information on key healthcare resource use beyond study mandated visits

Note: Subjects in Arm A will receive durvalumab plus gemcitabine/cisplatin combination therapy; subjects in Arm B will receive placebo plus gemcitabine/cisplatin therapy.

ctDNA Circulating tumor deoxyribonucleic acid; DCR Disease control rate; DoR Duration of response; EQ-5D-5L EuroQoL 5-dimension, 5-level health state utility index; HOSPAD Hospital resource use module; MSI Microsatellite instability; ORR Objective response rate; OS Overall survival; PFS Progression free survival; PGIS Patient Global Impression of Severity; PRO-CTCAE Patient reported outcomes-Common Terminology Criteria for Adverse Events; QLQ-BIL21 21-Item Cholangiocarcinoma and Gallbladder Cancer Quality of Life Questionnaire; RECIST Response Evaluation Criteria in Solid Tumors; TMB Tumor mutational burden.

1.2 Study design

This is a Phase III randomized, double-blind, placebo-controlled, multi-regional, international study to assess the efficacy and safety of first-line treatment with durvalumab in combination with gemcitabine/cisplatin versus placebo in combination with gemcitabine/cisplatin in subjects with previously untreated, unresectable locally advanced or metastatic BTC.

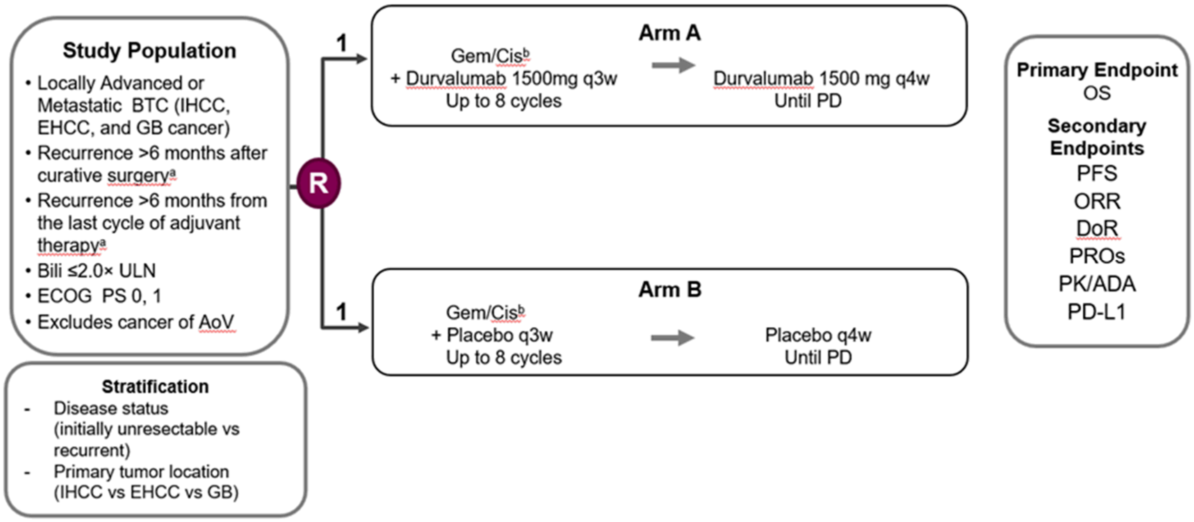
Subjects will be randomized 1:1 to receive one of the following treatments:

Arm A: Durvalumab plus gemcitabine/cisplatin combination therapy. Durvalumab 1500 mg via intravenous (IV) infusion every 3 weeks (q3w), starting on Cycle 1 in combination with cisplatin 25 mg/m^2 and gemcitabine 1000 mg/m^2 (each administered on Days 1 and 8 q3w) up to 8 cycles, followed by durvalumab 1500 mg as monotherapy every 4 weeks (q4w) until clinical progression or RECIST 1.1-defined radiological progression of disease (PD), unless there is unacceptable toxicity, withdrawal of consent or another discontinuation criterion is met (CSP Section 7.1 for additional details on discontinuation of study treatment). Durvalumab or placebo dose modification for subjects ≤ 30 kg is presented in Section 6.1.1.1. of the CSP.

Arm B: Placebo plus gemcitabine/cisplatin combination therapy. Placebo via IV infusion q3w, starting on Cycle 1 in combination with cisplatin 25 mg/m^2 and gemcitabine 1000 mg/m^2 (each administered on Days 1 and 8 q3w) up to 8 cycles, followed placebo monotherapy q4w until clinical progression or RECIST 1.1-defined radiological PD, unless there is unacceptable toxicity, withdrawal of consent or another discontinuation criterion is met (CSP Section 7.1 for additional details on discontinuation of study treatment).

An overview of the study design is shown in Figure 1.

Figure 1: Overview of study design



[a] Subjects with recurrence >6 months after curative surgery without adjuvant therapy or >6 months after adjuvant therapy will be included.
[b] Cisplatin (25 mg/m2) and gemcitabine (1000 mg/m2), each administered on Days 1 and 8, q3w for 8 cycles.
ADA Anti-drug antibody; AoV Ampulla of Vater; Bili Bilirubin; BTC Biliary tract cancer; DoR Duration of response; ECOG Eastern Cooperative Oncology Group; EHCC Extrahepatic cholangiocarcinoma; GB Gallbladder; Gem/Cis Gemcitabine plus cisplatin; IHCC Intrahepatic cholangiocarcinoma; ORR Objective

response rate; OS Overall survival; PD Progressive disease; PD-L1 Programmed cell death ligand-1; PFS Progression free survival; PK Pharmacokinetic; PRO Patient reported outcome; PS Performance status; q3w Every 3 weeks; q4w Every 4 weeks; ULN Upper limit of normal.

1.3 Number of subjects

Approximately 672 subjects (336 subjects per treatment arm) will be randomized 1:1 to durvalumab plus gemcitabine/cisplatin combination therapy or placebo plus gemcitabine/cisplatin combination therapy. The randomization will be stratified by disease status (initially unresectable versus recurrent) and primary tumor site (intrahepatic cholangiocarcinoma versus extrahepatic cholangiocarcinoma versus gallbladder cancer).

Approximately 130 Chinese subjects will be randomized in this study. The Sponsor will close the global study enrolment to all sites apart from sites in China at an appropriate time to ensure approximately 672 subjects are randomized to global study population. Recruitment of subjects from sites in China will continue until 130 Chinese subjects are randomized. Subjects randomized in China prior to the last subject randomized in the Global Cohort (which was on 18th Dec 2020) will be included in both Global Cohort and China Cohort. Enrolment of subjects from sites in China will be actively managed by the Sponsor to ensure there is no significant over recruitment of subjects from sites in China. Subjects randomized in China after last subject randomized in the global cohort will only be analyzed in the China Cohort. The analysis in China cohort will be performed when the OS data from the Chinese subjects is of similar maturity to those of the global cohort where significant clinical efficacy is established, e.g., if OS efficacy is established at the primary analysis, a similar maturity to this will be used for the consistency evaluation. China cohort includes patients from the Global Cohort randomized in China, and patients randomized in China after the last subject randomized in the Global Cohort.

This SAP describes only analyses of the Global Cohort. Analysis of the China cohort will be detailed in a separate SAP.

The study is powered to demonstrate superiority in the overall survival (OS) benefit of durvalumab plus gemcitabine/cisplatin (Arm A) versus placebo plus gemcitabine/cisplatin (Arm B) in subjects with previously untreated, unresectable locally advanced or metastatic BTC.

A hypothesis of improved OS will be tested when:

- Approximately 397 OS events have occurred across Arm A and Arm B (59% maturity) (Interim Analysis [IA-2]) and

- Approximately 496 OS events have occurred across Arm A and Arm B (74% maturity) (Final Analysis [FA]).

The primary analysis of OS is based on a log-rank test for the interim analysis and a FH (0, 1) test for the final analysis. The log-rank test will also be performed at the final analysis as a sensitivity analysis.

If the true average OS hazard ratio (HR) is 0.745, approximately 496 OS events will provide 90% power to demonstrate statistical significance at the 4.20% level (using a 2-sided significance level) at FA when using log-rank test based on O'Brien Fleming alpha spending function with 397 events at IA2, where 0.1% type I error is allocated to ORR analysis at IA1 and 4.9% is allocated to OS analysis. With a 21-month recruitment period and a minimum follow-up period of 19 months assumed, it is anticipated that this analysis will be performed approximately 40 months after the first subject is randomized.

With a log-rank test at IA-2 and a FH(0, 1) test at the final analysis, the overall power is at least 86% based on an assumed average HR of 0.745 under the assumption of proportional hazards or up to a 6-month delayed effect (i.e., delayed separation of the OS curves by up to 6 months).

Simulation studies were performed and demonstrated that the proposed method can control type I error and gain power compared to log-rank test under delayed effect scenarios. Simulations were performed and they demonstrated the proposed method can improve power compared to the log-rank test under scenarios of delayed effect. (Table 5).

Table 5: Power at IA and overall study under various scenarios using different tests

Scenario	**IA: log-rank** **FA: log-rank**	**IA: log-rank** **FA: FH(0, 1)**	**IA: FH(0, 1)** **FA: FH(0, 1) [2]**
No delay (Proportional Hazards) HR = 0.745	75% 90%	75% 86%	60% 80%
Delay 3 mo but average HR 0.745 at FA (HR = 0.685 after delayed period)	67% 89%	67% 93%	76% 93%
Delay 6 mo but average HR 0.745 at FA (HR = 0.609 after delayed period)	56% 89%	56% 97%	81% 97%
Delay 6 mo and HR = 0.64 after delayed period [1]	46% 81%	46% 93%	71% 92%
Delay 6 mo and HR = 0.67 after delayed period [1]	38% 73%	38% 86%	61% 88%

Note: [1] HR 0.64 and 0.67 were observed based on AZ Sponsored studies with durvalumab where a delay in treatment effect was observed. [2] This column is for reference only.
Note: Simulations were performed based on TOPAZ-1 study design: N = 672, 1:1 randomization ratio, control arm median OS 11.7 mo, enrollment 21 mo (A = 21), non-uniform enrollment pattern with weight w =1.5, i.e, by month t, the proportion of cumulative enrollment is $\left(\frac{t}{A}\right)^{w}$. IA and FA are analyzed when 397 events and 496 events are observed respectively. Alpha spending at IA is according to the current protocol of O'Brien-Fleming spending function based on log-rank test, 0.0119 (1-sided). Power for weighted log-rank tests is based on 10,000 simulations for each scenario.

Further simulations with more iterations were also performed to particularly study the type I error control at IA and overall study in Section 4.2.1. Various scenarios were considered in the simulations and they demonstrated strong control of type I error (Table 19, Table 20).

The study is considered to have met its primary objective if Arm A is statistically significantly superior to Arm B either at the time of the interim analysis for early testing for superiority in OS or at the final analysis.

2 ANALYSIS SETS

2.1 Definition of analysis sets

There are four analysis sets defined for this study. Definitions of the analysis sets for each outcome variable are provided in Table 6.

Table 6: Summary of outcome variables and analysis populations

Outcome variable	Analysis set
Efficacy Data	
OS	Full analysis set
PFS	Full analysis set
ORR*, DoR	Full analysis set
ORR, BoR and DoR	Full analysis set – subjects with >=32 weeks follow-up (for IA-1 analysis)
PRO endpoints	PRO analysis set
Study Population /Demography Data	
Demography characteristics	Full analysis set
Baseline and disease characteristics	Full analysis set
Important deviations	Full analysis set
Medical/surgical history	Full analysis set
Concomitant medications/procedures	Full analysis set
Biomarker Data	
Biomarker data	Full analysis set
PK Data	
PK data	PK analysis set
Safety Data	
Exposure	Safety analysis set
AEs	Safety analysis set
Laboratory measurements	Safety analysis set
Physical examinations and vital signs	Safety analysis set
ECGs	Safety analysis set
WHO/ECOG PS	Safety analysis set
ADA data	ADA analysis set

*Subjects who are evaluable for the analysis of ORR are those with measurable disease at baseline. Subjects who are evaluable for the analysis of DoR are those who responded in the ORR analysis.

AE Adverse event; DoR Duration of response; ECOG Eastern Cooperative Oncology Group; ORR Objective response rate; OS Overall survival; PS Performance status; PFS Progression free survival; PK Pharmacokinetics; PRO Patient reported outcome.

2.1.1 Full analysis set (Intention to treat (ITT))

The full analysis set (FAS) will include all randomized subjects. The FAS will be used for all efficacy analyses. Treatment arms will be compared on the basis of randomized study treatment, regardless of the treatment actually received. Subjects who were randomized but did not subsequently go on to receive study treatment are included in the analysis in the treatment arm to which they were randomized. The analysis of data using the FAS therefore follows the principles of ITT.

Analysis of ORR will be based on subjects in the FAS who have measurable disease at baseline (refer to Section 3.2.3). Analysis of DoR will be based on subjects in the FAS who achieve objective response (refer to Section 3.2.4).

For IA-1 an additional analysis set will be defined: FAS subjects with an opportunity for at least 32 weeks of follow up at the time of IA-1 DCO (FAS-32w, i.e., randomized ≥ 32 weeks prior to IA-1 data cut-off (DCO)).

2.1.2 PRO analysis set

For each patient reported outcome (PRO) questionnaire, a separate analysis set will be defined. The patient reported outcome analysis set will include all subjects from the FAS, except for subjects with no questionnaire translation available or who did not complete questionnaires due to physical limitations (e.g. blind), illiteracy, or other language reasons. All PRO analyses will take place using the PRO analysis set.

2.1.3 Safety analysis set

The safety analysis set (SAF) will consist of all subjects who received at least 1 dose of study treatment. Safety data will not be formally analyzed but summarized descriptively using the SAF, according to the treatment received. Erroneously treated subjects (e.g., those randomized to durvalumab but actually given placebo) will be summarized according to the treatment they actually received. If a subject only receives therapy from the placebo arm, they will be summarized in the placebo treatment group. If a subject receives any amount of durvalumab, they will be summarized in the durvalumab treatment group.

2.1.4 Pharmacokinetics analysis set

The pharmacokinetics (PK) analysis set includes all subjects who receive at least 1 dose of durvalumab per the protocol for whom any post-dose data are available. The population will be defined by the Study Physician, Clinical Pharmacologist or PK Scientist, and Statistician prior to any analyses being performed. The PK analysis set will be summarized according to the treatment actually received.

2.1.5 ADA analysis set

The anti-drug antibody (ADA) analysis set will include all subjects who have non-missing baseline ADA and at least 1 non-missing post-baseline ADA results. All major ADA analyses will be based on the ADA analysis set.

2.2 Violations and deviations

For this study, the following general categories will be considered important protocol deviations (IPDs) and will be programmatically derived from the electronic case report form (eCRF) data. These will be listed and summarized by randomized treatment group and discussed in the clinical study report (CSR) as appropriate:

- Subjects randomized but who did not receive study treatment (Deviation 1).
- Subjects who deviate from key entry criteria per the CSP (Deviation 2).
  - Inclusion 5: Histologically confirmed, unresectable advanced or metastatic adenocarcinoma of biliary tract, including cholangiocarcinoma (intrahepatic or extrahepatic) and gallbladder carcinoma.
  - Inclusion 6: Subjects with previously untreated disease if unresectable or metastatic at initial diagnosis will be eligible.
  - Inclusion 7: Subjects who developed recurrent disease >6 months after surgery with curative intent and, if given, >6 months after the completion of adjuvant therapy (chemotherapy and/or radiation) will be eligible.
  - Inclusion 8: WHO/ECOG PS 0 or 1 at enrolment.
  - Exclusion 1: Ampullary carcinoma.
- Baseline RECIST scan >42 days before randomization (based upon a 28-day screening period plus 2 weeks allowance, so that only serious violators are identified (Deviation 3).
- No baseline RECIST 1.1 assessment on or before date of randomization (Deviation 4).
- Received prohibited concomitant medications (including other anti-cancer agents) (Deviation 5). Please refer to the CSP Section 6.4 for those medications that are detailed as being ‘excluded’ from permitted use during the study. This will be used as a guiding principle for the physician review of all medications prior to database lock.

- Subjects randomized who received their randomized study treatment at an incorrect dose or received an alternative study treatment to that which they were randomized (Deviation 6).

- Subjects who developed discontinuation criteria during the study but were not discontinued (Deviation 7).

Subjects who receive the wrong treatment at any time will be included in the SAF as described in Section 2.1.3. During the study, decisions on how to handle errors in treatment dispensing (with regard to continuation/discontinuation of study treatment or, if applicable, analytically) will be made on an individual basis with written instruction from the study team leader and/or statistician.

The important protocol deviations will be listed and summarized by randomized treatment group. Deviation 1 (subjects randomized but who did not receive durvalumab or matching placebo) will lead to exclusion from the safety analysis set. None of the other deviations will lead to subjects being excluded from the analysis sets described in Section 2.1 (with the exception of the PK analysis set, if the deviation is considered to impact upon PK).

A per-protocol analysis excluding subjects with specific important protocol deviations is not planned, however, a “deviation bias” sensitivity analysis may be performed on the progression free survival endpoint excluding subjects with deviations that may affect the efficacy of the trial therapy if >10% of subjects in either treatment group:

- Did not have the intended disease or indication or

- Did not receive any randomized therapy.

The need for such a sensitivity analysis will be determined following review of the protocol deviations ahead of database lock and will be documented prior to the primary analysis being conducted.

In addition to the programmatic determination of the deviations above, other study deviations captured from the case report form (CRF) module for inclusion/exclusion criteria will be tabulated and listed. Any other deviations from monitoring notes or reports will be reported in an appendix to the CSR.

3 PRIMARY AND SECONDARY VARIABLES

3.1 Derivation of RECIST visit responses

For all subjects, the RECIST tumor response data will be used to determine each subject’s visit response according to RECIST version 1.1 (Appendix F of CSP). It will also be used to determine if and when a subject has progressed in accordance with RECIST and their best objective response (BoR) to study treatment.

Baseline radiological tumor assessments are to be performed no more than 28 days before the start of randomized treatment and ideally as close as possible to the start of study treatment. Tumor assessments are then performed every 6 weeks (± 1 week) for first 24 weeks (relative to date of randomization) then every 8 weeks (± 1 week) thereafter (relative to date of randomization) until RECIST 1.1 defined radiological disease progression plus at least 1 additional follow-up scan.

If an unscheduled assessment is performed, and the subject has not progressed, every attempt should be made to perform the subsequent assessments at their scheduled visits. This schedule is to be followed in order to minimize any unintentional bias caused by some subjects being assessed at a different frequency than other subjects.

From the investigator’s review of the imaging scans, the RECIST tumor response data will be used to determine each subject’s visit response according to RECIST version 1.1. At each visit, subjects will be programmatically assigned a RECIST 1.1 visit response of complete response (CR), partial response (PR), stable disease (SD) or progressive disease (PD) using the information from target lesions (TLs), non-target lesions (NTLs) and new lesions and depending on the status of their disease compared with baseline and previous assessments. If a subject has had a tumor assessment that cannot be evaluated then the subject will be assigned a visit response of not evaluable (NE), (unless there is evidence of progression in which case the response will be assigned as PD).

Refer to Section 3.1.1 for the definitions of CR, PR, SD and PD.

Subjects with measurable disease i.e. at least one target lesion at baseline, will be entered in this study (inclusion criteria 9). If a subject with non-measurable disease or no evidence of disease (NED assessed at baseline by computed tomography (CT) / magnetic resonance imaging (MRI) enters the study, RECIST will be modified to allow the assessment of progression due to new lesions in patients with no evidence of disease at baseline.

3.1.1 Target lesions (TLs)

Measurable disease is defined as having at least one measurable lesion, not previously irradiated, which is ≥10 mm in the longest diameter (LD), (except lymph nodes which must have short axis ≥15 mm) CT or MRI and which is suitable for accurate repeated measurements. A subject can have a maximum of five measurable lesions recorded at baseline with a maximum of two lesions per organ (representative of all lesions involved and suitable for accurate repeated measurement) and these are referred to as target lesions (TLs). If more than one baseline scan is recorded, then measurements from the one that is closest and prior to randomization will be used to define the baseline sum of TLs. It may be the case that, on occasion, the largest lesion does not lend itself to reproducible measurement. In which circumstance the next largest lesion, which can be measured reproducibly, should be selected.

All other lesions (or sites of disease) not recorded as TL should be identified as non-target lesions (NTLs) at baseline. Measurements are not required for these lesions, but their status should be followed at subsequent visits.

TL visit responses are described in Table 7 below.

Table 7: TL visit responses (RECIST 1.1)

Visit Responses	Description
Complete response (CR)	Disappearance of all TLs. Any pathological lymph nodes selected as TLs must have a reduction in short axis to <10mm.
Partial response (PR)	At least a 30% decrease in the sum of diameters of TLs, taking as reference the baseline sum of diameters as long as criteria for PD are not met.
Progressive disease (PD)	A ≥20% increase in the sum of diameters of TLs and an absolute increase of ≥5mm, taking as reference the smallest sum of diameters since treatment started including the baseline sum of diameters.
Stable disease (SD)	Neither sufficient shrinkage to qualify for PR nor sufficient increase to qualify for PD.
Not evaluable (NE)	Only relevant in certain situations (i.e. if any of the TLs were not assessed or not evaluable or had a lesion intervention at this visit; and scaling up could not be performed for lesions with interventions). Note: If the sum of diameters meets the progressive disease criteria, progressive disease overrides not evaluable as a TL response.

Visit Responses	Description
Not applicable (NA)	No TLs are recorded at baseline.

Rounding of TL data

For calculation of PD and PR for TLs percentage changes from baseline and previous minimum should be rounded to one decimal place (d.p.) before assigning a TL response. For example, 19.95% should be rounded to 20.0% but 19.94% should be rounded to 19.9%.

Missing TL data

For a visit to be evaluable then all TL measurements should be recorded. However, a visit response of PD should still be assigned if any of the following occurred:

- A new lesion is recorded.
- A NTL visit response of PD is recorded.
- The sum of TLs is sufficiently increased to result in a 20% increase, and an absolute increase of ≥5mm, from nadir even assuming the non-recorded TLs have disappeared.

Note: the nadir can only be taken from assessments where all the TLs had a LD recorded.

If there is at least one TL measurement missing and a visit response of PD cannot be assigned, the visit response is NE.

If all TL measurements are missing, then the TL visit response is NE. Overall visit response will also be NE, unless there is a progression of non-TLs or new lesions, in which case the response will be PD.

Lymph nodes

For lymph nodes, if the size reduces to <10mm then these are considered non-pathological. However, a size will still be given, and this size should still be used to determine the TL visit response as normal. In the special case where all lymph nodes are <10mm and all other TLs are 0mm then although the sum may be >0mm the calculation of TL response should be over-written as a CR.

TL visit responses subsequent to CR

Only CR, PD or NE can follow a CR. If a CR has occurred, then the following rules at the subsequent visits must be applied:

- Step 1: If all lesions meet the CR criteria (i.e. 0mm or <10mm for lymph nodes) then response will be set to CR irrespective of whether the criteria for PD of TL is also met i.e. if a lymph node LD increases by 20% but remains <10mm.

- Step 2: If some lesion measurements are missing but all other lesions meet the CR criteria (i.e. 0mm or <10mm for lymph nodes) then response will be set to NE irrespective of whether, when referencing the sum of TL diameters, the criteria for PD are also met.

- Step 3: If not all lesions meet the CR criteria (i.e. a pathological lymph node selected as TL has short axis ≥10mm and an absolute increase of ≥5mm, taking as reference the smallest short axis for the same TL since treatment started including the baseline or the reappearance of previously disappeared lesion) or a new lesion appears, then response will be set to PD.

- Step 4: If after steps 1 – 3 a response can still not be determined the response will be set to remain as CR.

TL too big to measure

If a TL becomes too big to measure this should be indicated in the database and a size (‘x’) above which it cannot be accurately measured should be recorded. If using a value of x in the calculation of TL response would not give an overall visit response of PD, then this will be flagged and reviewed by the study team blinded to treatment assignment. It is expected that a visit response of PD will remain in the vast majority of cases.

TL too small to measure

If a TL becomes too small to measure, then this will be indicated as such on the case report form and a value of 5mm will be entered into the database and used in TL calculations. However, a smaller value may be used if the radiologist has not indicated ‘too small to measure’ on the case report form and has entered a smaller value that can be reliably measured. If a TL response of PD results (at a subsequent visit) then this will be reviewed by the study team blinded to treatment assignment.

Irradiated lesions/lesion intervention

Previously irradiated lesions (i.e. lesion irradiated prior to entry into the study) should be recorded as NTLs and should not form part of the TL assessment.

Any TL (including lymph nodes), which has had intervention during the study (for example, irradiation / palliative surgery / embolization), should be handled in the following way. Once a lesion has had intervention then it should be treated as having intervention for the remainder of the study noting that an intervention will most likely shrink the size of tumors:

- Step 1: the diameters of the TLs (including the lesions that have had intervention) will be summed and the calculation will be performed in the usual manner. If the visit response is PD, this will remain as a valid response category.

- Step 2: If there was no evidence of progression after step 1, treat the lesion diameter (for those lesions with intervention) as missing and if ≤1/3 of the TLs have missing measurements then scale up as described in the 'Scaling' section below. If the scaling results in a visit response of PD then the subject would be assigned a TL response of PD.

- Step 3: If, after both steps, PD has not been assigned, then, if appropriate (i.e. if ≤ 1/3 of the TLs have missing measurements), the scaled sum of diameters calculated in step 2 should be used, and PR or SD then assigned as the visit response. Subjects with intervention are evaluable for CR as long as all non-intervened lesions are 0 (or <10mm for lymph nodes) and the lesions that have been subject to intervention have a value of 0 (or <10mm for lymph nodes) recorded. If scaling up is not appropriate due to too few non-missing measurements, then the visit response will be set as NE.

At subsequent visits, the above steps will be repeated to determine the TL and overall visit response. When calculating the previous minimum, lesions with intervention should be treated as missing and scaled up (as per step 2 above).

Scaling (applicable only for irradiated lesions/lesion intervention)

If >1/3 of TL measurements are missing (because of intervention) then the TL response will be NE, unless the sum of diameters of non-missing TL would result in PD (i.e. if using a value of 0 for missing lesions, the sum of diameters has still increased by 20% or more compared to nadir and the sum of TLs has increased by ≥5mm from nadir).

If ≤1/3 of the TL measurements are missing (because of intervention) then the results will be scaled up (based on the sizes at the nadir visit to give an estimated sum of diameters) and this will be used in calculations; this is equivalent to comparing the visit sum of diameters of the non-missing lesions to the nadir sum of diameters excluding the lesions with missing measurements.

Example of scaling

Lesion 5 is missing at the follow-up visit; the nadir TL sum including lesions 1-5 was 74mm.

The sum of lesions 1-4 at the follow-up is 68mm. The sum of the corresponding lesions at the nadir visit is 62mm.

Scale up as follows to give an estimated TL sum of 81mm:

68 x 74 / 62 = 81mm

CR will not be allowed as a TL response for visits where there is missing data. Only PR, SD or PD (or NE) could be assigned as the TL visit response in these cases. However, for visits with ≤1/3 lesion assessments not recorded, the scaled-up sum of TLs diameters will be included when defining the nadir value for the assessment of progression.

Lesions that split in two

If a TL splits in two, then the LDs of the split lesions should be summed and reported as the LD for the lesion that split.

Lesions that merge

If two TLs merge, then the LD of the merged lesion should be recorded for one of the TL sizes and the other TL size should be recorded as 0cm.

Change in method of assessment of TLs

CT and MRI are the only methods of assessment that can be used within this trial. If a change in method of assessment occurs, between CT and MRI this will be considered acceptable and no adjustment within the programming is needed.

3.1.2 Non-target lesions (NTLs) and new lesions

At each visit, the investigator should record an overall assessment of the NTL response. This section provides the definitions of the criteria used to determine and record overall response for NTL at the investigational site at each visit.

NTL response will be derived based on the investigator's overall assessment of NTLs as shown in Table 8:

Table 8: NTL visit responses

Visit Responses	Description
Complete response (CR)	Disappearance of all NTLs present at baseline with all lymph nodes non-pathological in size (<10 mm short axis).
Progressive disease (PD)	Unequivocal progression of existing NTLs. Unequivocal progression may be due to an important progression in one lesion only or in several lesions. In all cases, the progression MUST be clinically significant for the physician to consider changing (or stopping) therapy.
Non-CR/Non-PD	Persistence of one or more NTLs with no evidence of progression.

Visit Responses	Description
Not evaluable (NE)	Only relevant when one or some of the NTLs were not assessed and, in the investigator's opinion, they are not able to provide an evaluable overall NTL assessment at this visit. Note: For subjects without TLs at baseline, this is relevant if any of the NTLs were not assessed at this visit and the progression criteria have not been met.
Not applicable (NA)	Only relevant if there are no NTLs at baseline.

To achieve ‘unequivocal progression’ on the basis of NTLs, there must be an overall level of substantial worsening in non-target disease such that, even in the presence of SD or PR in TLs, the overall tumor burden has increased sufficiently to merit a determination of disease progression. A modest ‘increase’ in the size of one or more NTLs is usually not sufficient to qualify for unequivocal progression status.

Details of any new lesions will also be recorded with the date of assessment. The presence of one or more new lesions is assessed as progression.

A lesion identified at a follow up assessment in an anatomical location that was not scanned at baseline is considered a new lesion and will indicate disease progression.

The finding of a new lesion should be unequivocal: i.e. not attributable to differences in scanning technique, change in imaging modality or findings thought to represent something other than tumor.

New lesions will be identified via a Yes/No tick box. The absence and presence of new lesions at each visit should be listed alongside the TL and NTL visit responses.

A new lesion indicates progression so the overall visit response will be PD irrespective of the TL and NTL response.

Symptomatic progression is not a descriptor for progression of NTLs: it is a reason for stopping study therapy and will not be included in any assessment of NTLs.

Subjects with ‘symptomatic progression’ requiring discontinuation of treatment without objective evidence of disease progression at that time should continue to undergo tumor assessments where possible until objective disease progression is observed.

3.1.3 Overall RECIST 1.1 visit response

Table 9 defines how the previously defined TL and NTL visit responses will be combined with new lesion information to give an overall visit response.

Table 9: Overall visit responses

Target	Non-target	New lesions	Overall visit response
CR	CR or NA	No (or NE)	CR
CR	Non-CR/Non-PD or NE	No (or NE)	PR
PR	Non-PD or NE or NA	No (or NE)	PR
SD	Non-PD or NE or NA	No (or NE)	SD
PD	Any	Any	PD
Any	PD	Any	PD
Any	Any	Yes	PD
NE	Non-PD or NE or NA	No (or NE)	NE
NA	CR	No (or NE)	CR
NA	Non-CR/Non-PD	No (or NE)	SD
NA	NE	No (or NE)	NE
NA	NA	No (or NE)	NED

CR Complete response; NA Not Applicable; NE Not evaluable; PD Progressive disease; PR Partial response; SD Stable disease; NED No evidence of disease

3.1.4 Independent review

A planned blinded independent central review (BICR) of radiological imaging data will be carried out using RECIST version 1.1. All radiological scans for all subjects (including those at unscheduled visits, or outside visit windows) will be collected on an ongoing basis and sent to an AstraZeneca appointed Contract Research Organisation (CRO) for central analysis. The imaging scans will be reviewed by two independent radiologists using RECIST 1.1 and will be adjudicated, if required (i.e., two reviewers' review the scans and adjudication is performed by a separate reviewer in case of a disagreement). For each subject, the BICR will define the overall visit response (i.e., the response obtained overall at each visit by assessing TLs, NTLs and new lesions) data and no programmatic derivation of visit response is necessary (for subjects with TLs at baseline: CR, PR, SD, PD, NE; for subjects with NTLs only: CR, SD, PD, NE). If a subject has had a tumor assessment that cannot be evaluated, then the subject will be assigned a visit response of NE (unless there is evidence of progression in which case the response will be assigned as PD). Tumor assessments/scans contributing towards a particular visit may be performed on different dates and for the central review the date of progression for each reviewer will be provided based on the earliest of the scan dates of the component that triggered the progression.

If adjudication is performed, the reviewer that the adjudicator agreed with will be selected as a single reviewer (note in the case of more than one review period, the latest adjudicator decision will be used). In the absence of adjudication, the records for all visits for a single reviewer will be used. The reviewer selected in the absence of adjudication will be the reviewer who read the baseline scan first. The records from the single selected reviewer will be used to report all BICR RECIST information including dates of progression, visit response, censoring and changes in target lesion dimensions. Endpoints (of ORR, PFS and DoR) will be derived programmatically from this information.

Results of this independent review will not be communicated to investigators and the management of subjects will be based solely upon the results of the RECIST 1.1 assessment and physical examination conducted by the investigator.

A BICR will be performed for the first IA-1 database lock for ORR/DoR, which will cover all scheduled and unscheduled on-protocol scans obtained up to data cut-off (DCO) for all randomized subjects who have had the opportunity to be followed for at least 32 weeks (± 1 week). This will be to derive ORR and DoR for IA-1. All other images will be collected and stored for potential BICR if considered necessary by AstraZeneca.

Further details of the BICR will be documented in the Independent Review Charter (IRC).

3.2 Efficacy variables

3.2.1 Overall survival (OS)

The primary endpoint of the trial is overall survival defined as time from the date of randomization until death due to any cause regardless of whether the subject withdraws from randomized therapy or receives another anti-cancer therapy (i.e. date of death or censoring – date of randomization + 1). Any subject not known to have died at the time of analysis will be censored based on the last recorded date on which the subject was known to be alive (maximum SUR_DAT, recorded within the SURVIVE module of the eCRF or the latest eCRF date).

Note: Survival calls will be made following the date of DCO for each analysis (these contacts should generally occur within 7 days of the DCO). If subjects are confirmed to be alive or if the death date is post the DCO date, these subjects will be censored at the date of DCO. The status of ongoing, withdrawn (from the study) and "lost to follow-up" subjects at the time of the final OS analysis should be obtained by the site personnel by checking the subject's notes, hospital records, contacting the subject's general practitioner and checking publicly-available death registries. In the event that the subject has actively withdrawn consent to the processing of their personal data, the vital status of the subject can be obtained by site personnel from publicly available resources where it is possible to do so under applicable local laws.

Note: For any OS analysis performed prior to the final OS analysis, in the absence of survival calls being made, it may be necessary to use all relevant CRF fields to determine the last recorded date on which the subject was known to be alive for those subjects still on treatment (since the SURVIVE module is only completed for subjects off treatment if a survival sweep is not performed). The last date for each individual subject is defined as the latest among the following dates recorded on the case report forms (CRFs):

- AE start and stop dates
- Admission and discharge dates of hospitalization
- Study treatment date
- End of treatment date
- Laboratory test dates
- Date of vital signs

• Disease assessment dates on RECIST CRF

• Start and stop dates of subsequent anticancer therapy

• Date last known alive on survival status CRF

• End of study date

If a subject is known to have died where only a partial death date is available, then the date of death will be imputed as the latest of the last date known to be alive + 1 from the database and the death date using the available information provided

a. For Missing day only – using the 1st of the month
b. For Missing day and Month – using the 1st of January

If there is evidence of death but the date is entirely missing, it will be treated as missing, i.e. censored at the last known alive date.

3.2.2 Progression free survival (PFS)

The secondary endpoint PFS (per RECIST 1.1 as assessed by the site Investigator) will be defined as the time from the date of randomization until the date of RECIST 1.1-defined radiological PD or death (by any cause in the absence of progression) regardless of whether the subject withdraws from therapy or receives another anticancer therapy prior to progression (i.e., date of PFS event or censoring – date of randomization + 1). Subjects who have not progressed or died at the time of analysis will be censored at the time of the latest date of assessment from their last evaluable RECIST 1.1 assessment. However, if the subject progresses or dies after 2 or more missed visits, the subject will be censored at the time of the latest evaluable RECIST 1.1 assessment prior to the 2 missed visits (Note: NE visit is not considered as missed visit).

Given the scheduled visit assessment scheme (i.e. every 6 weeks for the first 24 weeks then every 8 weeks thereafter) the definition of 2 missed visits will change as follows:

- If the subject has no evaluable visits (no other visits than with the overall visit response of “NE”) or does not have baseline data, they will be censored at Day 1 unless they die within 2 visits of baseline (12 weeks plus 1 week allowing for a late assessment within the visit window), then they will be treated as an event with date of death as the event date.

- If the previous RECIST assessment is day 1 then two missing visits will equate to 13 weeks since the previous RECIST assessment, allowing for a late visit (i.e. 2 x 6 weeks +

1 week for a late assessment = 13 weeks).

- If the previous RECIST assessment is greater than day 1 and less than or equal to study day 119 (i.e. week 17) then two missing visits will equate to 14 weeks since the previous RECIST assessment, allowing for early and late visits (i.e. 2 x 6 weeks + 1 week for an early assessment + 1 week for a late assessment = 14 weeks).

- If the two missed visits occur over the period when the scheduled frequency of RECIST assessments changes from six-weekly to eight-weekly this will equate to 16 weeks (i.e. take the average of 6 and 8 weeks which gives 7 weeks and then apply same rationale, hence 2 x 7 weeks + 1 week for an early assessment + 1 week for a late assessment = 16 weeks). The time period for the previous RECIST assessment will be from study days 120 to 161 (i.e. week 17 to week 23).

- From week 23 (day 162) onwards (when the scheduling changes to eight-weekly assessments), two missing visits will equate to 18 weeks (i.e. 2 x 8 weeks + 1 week for an early assessment + 1 week for a late assessment = 18 weeks).

The PFS time will always be derived based on scan/assessment dates and not on visit dates.

RECIST 1.1 assessments/scans contributing toward a particular visit may be performed on different dates. The following rules will be applied:

- For Investigator assessments, the date of progression will be determined based on the earliest of the RECIST assessment/scan dates of the component that indicates progression.

- When censoring a subject for PFS, the subject will be censored at the latest of the scan dates contributing to a particular overall visit assessment.

Note: for TLs only the latest scan date is recorded out of all scans performed at that assessment for the TLs and similarly for NTLs only the latest scan date is recorded out of all scans performed at that assessment for the NTLs.

3.2.3 Objective response rate (ORR)

The secondary endpoint ORR is defined as the percentage of subjects with at least one investigator-assessed visit response of CR or PR and will be based on a subset of all randomized subjects with measurable disease at baseline per the site investigator. ORR will also be defined using the BICR data (at IA-1 only) to define a visit response of CR or PR

(Table 9), with the denominator defined as subset of all randomized subjects with measurable disease at baseline per BICR. ORR will also be analyzed in FAS as a sensitivity analysis.

Data obtained up until progression, or last evaluable assessment in the absence of progression, will be included in the assessment of ORR. Subjects who discontinue randomized treatment without progression, receive a subsequent anti-cancer therapy, and then respond will not be included as responders in the ORR.

3.2.4 Duration of response (DoR)

The secondary endpoint DoR (per RECIST 1.1 using Investigator assessment) will be defined as the time from the date of first documented response until date of documented progression or death in the absence of disease progression (i.e. date of PFS event or censoring – date of first response + 1). The end of response should coincide with the date of progression or death from any cause used for the PFS endpoint. The time of the initial response will be defined as the latest of the dates contributing towards the first visit response of PR or CR as defined by Table 9. If a subject does not progress following a response, then their DoR will use the PFS censoring time.

At IA-1, DoR will also be defined from BICR data.

3.2.5 Best objective response (BoR)

Best objective response (BoR) is calculated based on the overall visit responses from each tumor assessment, described in Section 3.1.3. It is the best response a subject has had following randomization, but prior to starting any subsequent cancer therapy and up to and including RECIST progression or the last evaluable assessment in the absence of RECIST progression. Categorization of BoR will be based on RECIST using the following response categories (in order from the best one to worst one): CR, PR, SD, NED (applies only to those subjects entering the study with no disease at baseline), PD and NE.

For determination of a best response of SD, the earliest of the dates contributing towards a particular overall visit assessment will be used. BoR or SD and NED should be recorded at least 6 weeks minus 1 week, i.e. at least 35 days (to allow for an early assessment within the assessment window), after randomization. For CR/PR, the initial overall visit assessment that showed a response will use the latest of the dates contributing towards a particular overall visit assessment.

BoR will be determined programmatically based on RECIST from the overall visit response using all BICR data up until the first progression event. It will also be determined programmatically based on RECIST using all site investigator data up until the first progression event. The denominators for each case will be consistent with those used in the ORR analysis.

For subjects whose progression event is death, BoR will be calculated based upon all evaluable RECIST assessments prior to death.

For subjects who die with no evaluable RECIST assessments, if the death occurs ≤13 weeks (i.e. 12 weeks + 1 week to allow for a late assessment within the assessment window) after randomization, then BoR will be assigned to the progression (PD) category. For subjects who die with no evaluable RECIST assessments, if the death occurs >13 weeks after randomization then BoR will be assigned to the NE category.

A subject will be classified as a responder if the RECIST criteria for a CR or PR, outlined in Table 9, are satisfied at any time following randomization, prior to RECIST progression and prior to starting any subsequent cancer therapy.

BoR will be defined from investigator assessments and additionally at IA-1 from BICR data.

3.2.6 Disease control rate (DCR)

DCR (per RECIST 1.1 as assessed by the Investigator) is defined as the rate of best objective response of NED, CR, PR, or SD according to RECIST 1.1. The denominator for DCR calculation is the number of subjects in the FAS.

DCR-24w is defined as the percentage of subjects who have a best objective response of NED, CR or PR (by week 24 + 7 days) or who have SD for at least 24 weeks (-7 days), following the date of randomization.

DCR-32w is defined as the percentage of subjects who have a best objective response of NED, CR or PR (by week 32 + 7 days) or who have SD for at least 32 weeks (-7 days), following the date of randomization.

DCR-48w is defined as the percentage of subjects who have a best objective response of NED, CR or PR (by week 48 + 7 days) or who have SD for at least 48 weeks (-7 days), following the date of randomization.

For IA-1 DCR will also be assessed by BICR assessments.

3.2.7 Change in tumor size

For supportive purposes percentage change from baseline in tumor size will be derived at each scheduled tumor assessment visit (hereafter referred to as week X for convenience). Best percentage change from baseline in tumor size will also be derived as the biggest decrease or, if no decrease, as the smallest increase in tumor size from baseline in the absence of a reduction and will include all assessments up to and including any evidence of progression (or prior to death in the absence of progression) prior to the start of subsequent anti-cancer therapy. Otherwise the last evaluable RECIST assessment if the subject has not died, progressed or started subsequent anti-cancer therapy.

This is based on RECIST 1.1 target lesion (TL) measurements taken at baseline and at the timepoint. Tumor size is the sum of the longest diameters of the TLs. TLs are measurable tumor lesions. Baseline for RECIST is defined to be the last evaluable assessment prior to randomization. The percentage change in TL tumor size at week X will be obtained for each subject taking the difference between the sum of the TLs at week X and the sum of the target lesions at baseline divided by the sum of the TLs at baseline times 100 (i.e. (week X - baseline) / baseline * 100).

Apply a window around the week X visit: Whenever tumor size data for the week X visit (Note: or visit at which progression was documented if before week X) is available then this should be used in the analysis. A windowing rule will be applied and will follow the protocol allowed visit window; therefore, any RECIST scan performed within ± 1 week of the protocol scheduled visit will be used for that visit, see Appendix B Table 27 for RECIST visit windowing.

If best percentage change cannot be calculated due to missing data (including if the subject has no TLs at baseline), a value of +20% will be imputed as the best percentage change from baseline in the following situations (otherwise best percentage change will be left as missing):

- If a subject has no post-baseline assessment and has died
- If a subject has new lesions or progression of NTLs or TLs
- If a subject has withdrawn due to PD and has no evaluable TL data before or at PD

Summaries for tumor size will be produced for Investigator assessments per RECIST 1.1. For IA-1 these will be calculated in FAS-32w for both Investigator and BICR assessments, with BICR assessments being of primary interest.

3.3 Patient reported outcome (PRO) variables

The following PRO questionnaires will be used to assess the patient experience, including global health status/health-related quality of life (HRQoL), functioning and symptoms: EORTC QLQ-C30 with the EORTC QLQ-BIL21 BTC disease specific module, PGIS, PRO-CTCAE and EQ-5D-5L. All items/questionnaires will be scored according to published guidelines or the developer's guidelines, if published guidelines are not available as described in the sections below. All PRO analyses will be based on the PRO analysis set, unless stated otherwise.

The PRO evaluations will be separated by on-treatment assessments (those taken on or before last dose of study treatment) and follow-up assessments (those taken after last dose of study treatment).

Descriptive summaries for absolute changes from baseline and summaries of response by visit will be reported for all on-treatment visits and follow-up visits month 1, month 2 and month 3.

Formal analysis of change from baseline using a mixed model repeated measures (MMRM) will only include on-treatment visits.

All available on-treatment and off-treatment PRO assessments will be used to determine best response, improvement based on best response and time to deterioration.

3.3.1 EORTC QLQ-C30

The EORTC QLQ-C30 consists of 30 questions that can be combined to produce 5 functional scales (physical, role, cognitive, emotional, and social), 3 symptom scales (fatigue, pain, and nausea/vomiting), and global health status/QoL scale. The EORTC QLQ-C30 will be scored according to the EORTC QLQ-C30 Scoring Manual (Fayers et al. 2001). An outcome variable consisting of a score from 0 to 100 will be derived for each of the symptom scales, each of the functional scales, and the global measure of health status scale in the EORTC QLQ-C30 according to the EORTC QLQ-C30 Scoring Manual. Higher scores on the global measure of health status and functional scales indicate better health status/function, but higher scores on symptom scales represent greater symptom severity. The EORTC QLQ-C30 functional and symptom scales, individual symptom items and global health status are derived as follows:

1. Calculate the average of the items that contribute to the scale or take the value of an individual item, i.e. the raw score (RS):

   $$RS = (I_1 + I_2 + \ldots + I_n) / n,$$

   where $I_1 + I_2 + \ldots + I_n$ are the items included in a scale and n is the number of items in a scale.

2. Use a linear transformation to standardize the raw score, so that scores range from 0 to 100, where a higher score represents a higher ("better") level of functioning, or a higher ("worse") level of symptoms.

   Functional scales: $\text{Score} = (1 - [RS - 1] / \text{range}) * 100$

Symptom scales/items; global health status: Score = ([RS – 1] / range) * 100,

where range is the difference between the maximum and the minimum possible value of RS.

The number of items and item range for each scale/item are displayed in Table 10 below.

Table 10: EORTC QLQ-C30 scales and scores

Scale/ item	Scale/ item abbreviation	Number of items (n)	Item range	Item numbers
Global health status/ QoL	QL	2	6	29, 30
Functional scales				
Physical	PF	5	3	1-5
Role	RF	2	3	6, 7
Cognitive	CF	2	3	20, 25
Emotional	EF	4	3	21-24
Social	SF	2	3	26, 27
Symptom scales				
Fatigue	FA	3	3	10, 12, 18
Pain	PA	2	3	9, 19
Nausea/ vomiting	NV	2	3	14, 15
Symptom items				
Dyspnoea	DY	1	3	8
Insomnia	SL	1	3	11
Appetite loss	AP	1	3	13
Constipation	CO	1	3	16
Diarrhoea	DI	1	3	17

EORTC European Organisation for Research and Treatment of Cancer; QLQ-C30 30-item core quality-of-life questionnaire.

For each subscale, if <50% of the subscale items are missing, then the subscale score will be divided by the number of non-missing items and multiplied by the total number of items on the subscales (Fayers et al. 2001). If at least 50% of the items are missing, then that subscale will be treated as missing. Missing single items are treated as missing. The reason for any missing questionnaire will be identified and recorded.

Definition of clinically meaningful changes - visit response and best overall response

Changes in score with baseline will be evaluated. A clinically meaningful change is defined as an absolute change in the score from baseline of ≥10 for scales from the EORTC QLQ-C30 (Osoba et al. 1998). For example, a clinically meaningful improvement in physical function (as assessed by EORTC QLQ-C30) is defined as an increase in the score from baseline of ≥10, whereas a clinically meaningful deterioration is defined as a decrease in the score from baseline of ≥10. At each post-baseline assessment, the change in global health status/QoL, symptoms, and functioning score from baseline will be categorized as improvement, no change, or deterioration as shown in Table 11.

Table 11: Mean change and clinically meaningful change - EORTC QLQ-C30

Score	Change from baseline	Visit response
EORTC QLQ-C30 global quality-of-life score	Increase of at least 10	Improvement
	Decrease of at least 10 or “Subject too sick to complete the questionnaires (disease under investigation)”	Deterioration
	Otherwise	No change
EORTC QLQ-C30 symptom score	Increase of at least 10 or “Subject too sick to complete the questionnaires (disease under investigation)”	Deterioration
	Decrease of at least 10	Improvement
	Otherwise	No change

Score	Change from baseline	Visit response
EORTC QLQ-C30 functional scales score	Increase of at least 10	Improvement
	Decrease of at least 10 or "Subject too sick to complete the questionnaires (disease under investigation)"	Deterioration
	Otherwise	No change

EORTC European Organisation for Research and Treatment of Cancer; QLQ-C30 30-Item Core Quality of Life Questionnaire.

A subject's best overall response in symptoms, function, or global health status/QoL will be derived as the best response the subject achieved based on evaluable PRO data collected during the study period including all on-treatment and off-treatment visits. The criteria in Table 12 will be used to assign a best response in symptoms, function, or global health status/QoL.

Table 12: Best response in EORTC QLQ-C30 and EORTC QLQ-BIL21 scores

Overall score response	Criteria
Missing	Subject has no evaluable baseline or post-baseline PRO assessment
Improved	Subject meets one of the following criteria: 1. Has 2 consecutive visit responses of "improvement" at least 14 days apart 2. Has 1 visit response of "improvement" with no further assessments and did not die within 2 PRO assessment visits
No change	Subject does not qualify for an overall score response of "improved" and meets one of the following criteria: 1. Has 2 consecutive visit responses of "no change" at least 14 days apart 2. Has 1 visit response of "no change" with no further assessments and did not die within 2 PRO assessment visits

Deterioration	Subject does not qualify for an overall score response of “improved” or “no change” and meets one of the following criteria: 1. Has 2 consecutive visit responses of “deterioration” at least 14 days apart 2. Has 1 visit response of “deterioration” and no further assessments 3. Has 1 visit response of “improvement” or “no change” followed by death within 2 PRO assessment visits

EORTC European Organisation for Research and Treatment of Cancer;
PRO Patient reported outcome; QLQ-C30 30-Item Core Quality of Life Questionnaire; QLQ-BIL21 21-Item Cholangiocarcinoma and Gallbladder Cancer Quality of Life Questionnaire.

Time to global health status/QoL function or symptom deterioration

Time to global health status/QoL, function or symptom deterioration will be defined as the time from the date of randomization until the date of the first clinically meaningful deterioration (as defined in Table 11) that is confirmed at a subsequent visit (except if it was the subject’s last available assessment) or death (by any cause) in the absence of a clinically meaningful deterioration, regardless of whether the subject discontinues the study treatment(s) or receives another anticancer therapy prior to global health status/QoL, function or symptom deterioration. Death will be included as an event only if it occurs within 2 PRO assessment visits from the last available PRO assessment.

Subjects whose global health status/QoL, function or symptoms (as measured by EORTC QLQ-C30) have not shown a clinically meaningful deterioration and who are alive at the time of the analysis will be censored at the time of their last PRO assessment, where the global health status/QoL, function, or symptom could be evaluated. Also, if global health status/QoL, function or symptoms deteriorates or the subject dies after 2 or more missed PRO assessment visits, the subject will be censored at the time of the last PRO assessment, where global health status/QoL or function could be evaluated prior to the 2 missed visits.

A sensitivity analysis of time to global health status/QoL, function or symptom deterioration will be conducted in which subject did not experience a clinically meaningful deterioration and deceased by the time of analysis will be censored at the last PRO assessment where the symptom could be evaluated or date of randomization if symptoms could not be evaluated as shown in Table 14.

To determine whether a PRO event should be censored due to extensive time between assessments (2 missed visits), examine the date of the last evaluable PRO assessment prior to the deterioration. In general, the elapsed time to the previous evaluable assessment should not be more than 2*(the protocol time between assessments) + 2*(the protocol allowed visit window of 3 days). If assessments are missed immediately after baseline, only 3 days for visit window is included, as there is no requirement to allow for an early visit.

Given the scheduled visit assessment scheme for TOPAZ (i.e. q3w ± 3 days from randomization for first 24 weeks then q4w ± 3 until last dose then two missed visits are defined as in Table 13

Table 13: EORTC QLQ-C30 and EORTC QLQ-BIL21 2 missed visit rules

Protocol Scheduled assessment for EORTC QLQ-C30/BIL21	**Previous non-missing post baseline assessment measured in the following window**	**Two missed visit window**
Q3w +/-3 days (up to Week 24)	Baseline only (Day 1 to on or prior to 1st dose)	2*3 weeks + 3 days (late visit). 45 days
	Up to pre Week 21 (1st dose – Day 144)	2*3 weeks + 3 days (early visit) + 3 days (late visit). 48 days
	Week 21 – pre Week 24 (Day 145 – Day 165)	3 weeks + 4 weeks + 3 days for early visit + 3 days late visit 55 days
Q4w +/-3 days (after Week 24)	Week 24 thereafter until last dose (Day 166 – last dose)	2*4 weeks + 3 days (early visit) and + 3 days (late visit) 62 days
Follow-up 30 Days +/- 3 Days (after last dose)	Last dose + 1 until 30 Day follow-up (Last dose + 1 – 30 Day follow-up visit)	1 month +30 days + 3 days (early visit) + 1 week (late visit) 70 days

Follow-up 1 month +/- 1 week (after 30-day follow-up)	Up to pre-3-month follow-up	2*1 month + 1 week (early visit) + 1 week (last visit) 74 days
	3-month follow-up until pre-4month follow-up	1 month + 2 month + 1 week (early visit) + 1 week (late visit) 104 days
Follow-up every 2 months +/- 1 week (after 4-month follow-up)	Up to pre-8-month follow-up	2*2 months + 1 week (early visit) + 1 week (late visit) 134 days
	8-month follow-up until pre-10month follow-up	2*2 months + 1 week (early visit) + 2 weeks (late visit) 141 days
	10-month follow-up until pre-12month follow-up	2 months + 6 months + 2 week (early visit) + 2 week (late visit) 268 days
Follow-up every 6 months +/- 2 weeks (after 12 months follow-up)	12-month follow-up until last follow-up	2*6 months + 2 weeks (early visit) + 2 weeks (late visit) 388 days

EORTC European Organisation for Research and Treatment of Cancer; QLQ-C30 30-Item Core Quality of Life Questionnaire; QLQ-BIL21 21-Item Cholangiocarcinoma and Gallbladder Cancer Quality of Life Questionnaire; Q3w Every 3 weeks; Q4W Every 4 weeks.

The population for the analysis of time to global health status/QoL or function deterioration will include a subset of the PRO analysis set who have baseline scores of ≥10.The population for the analysis of time to symptom deterioration will consist of a subset of the PRO analysis set subjects who have a baseline symptom score ≤90.

Table 14: Event and censoring rules for time to deterioration (symptoms, function, global health status/QoL)

Status	Date of Censoring for Main Analysis	Date of Censoring for Sensitivity Analysis
Clinically meaningful deterioration (at last visit or confirmed at a subsequent visit) prior to 2 or more missed PRO visits	Not censored (event at date of first assessment meeting criteria)	Not censored (event at date of first assessment meeting criteria)
No clinically meaningful deterioration and death (by any cause) at time of analysis prior to 2 or more missed PRO visits	Not censored (event at date of death)	Date of last evaluable PRO assessment or date of randomization if no evaluable baseline
No evaluable baseline or no evaluable post baseline PRO assessment and subject did not die within two visits of randomization (2 × 3 weeks + 3 days = 45 days)	Date of randomization	Date of randomization
No clinically meaningful deterioration and alive at time of analysis	Date of last evaluable PRO assessment	Date of last evaluable PRO assessment
Clinically meaningful deterioration or death after 2 or more missed PRO visits	Date of last evaluable PRO assessment prior to the 2 missed visits	Date of last evaluable PRO assessment prior to the 2 missed visits

PRO Patient reported outcome; QoL Quality of life.

Symptom improvement rate

Responses in symptoms for each visit (improvement, deterioration, and no change) based on Table 11 as well as the best overall response will be presented by treatment group. The symptom improvement rate will be defined as the number (%) of subjects with a best overall score response of “improved” in symptoms.

The denominator will consist of a subset of the PRO analysis set who have a baseline symptom score ≥10.

Global health status/QoL or function improvement rate

The global health status/QoL or function improvement rate will be defined as the number (%) of subjects with a best overall response of “improved” in QoL or function. The denominator will consist of a subset of the PRO analysis set who have a baseline global health status/QoL or function score ≤90.

3.3.2 EORTC QLQ-BIL21

The QLQ-BIL21 is a BTC-specific module from the EORTC comprising 21 questions to assess BTC symptoms. QLQ-BIL21 will be scored as described in Section 8 Appendices

EORTC QLQ – BIL 21 Scoring Procedure. The module includes 5 multi-item domain scales and 3 single-item scales. For all items and scales, high scores indicate increased symptomatology/more problems.

The scoring approach for the QLQ-BIL21 is identical in principle to that for the symptom scales/single items of the EORTC QLQ-C30. Similar to the symptom scales of the EORTC QLQ-C30, higher scores represent greater symptom severity.

Definition of clinically meaningful change - visit response and best overall response

Changes in score compared with baseline will be evaluated. A clinically meaningful change is defined as an absolute change in the score from baseline of ≥10 for scales/items from QLQ-BIL21. For example, a clinically meaningful deterioration or worsening in pain (as assessed by QLQ-BIL21) is defined as an increase in the score from baseline of ≥10. At each post-baseline assessment, the change in symptom score from baseline will be categorized as improved, no change, or deterioration, as shown in Table 15. A subject's best overall response in symptoms will be derived as the best response the subject achieved based on evaluable PRO data collected during the study period. The criteria in Table 12 will be used to assign a best response in symptom score.

Table 15: Mean change and clinically meaningful change - EORTC QLQ-BIL21

Score	Change from baseline	Visit response
QLQ-BIL21 symptom scales and items	Increase of at least 10 or "Subject too sick to complete the questionnaires (disease under investigation)"	Deterioration
	Decrease of at least 10	Improved
	Otherwise	No change

EORTC European Organisation for Research and Treatment of Cancer; QLQ-BIL21 21-Item Cholangiocarcinoma and Gallbladder Cancer Quality of Life Questionnaire.

Time to symptom deterioration

For each of the symptom scales/items in the QLQ-BIL21, time to symptom deterioration will be defined as the time from randomization until the date of the first clinically meaningful symptom deterioration that is confirmed at a subsequent visit (except if it was the subject's last available assessment) or death (by any cause) in the absence of a clinically meaningful symptom deterioration, regardless of whether the subject discontinues the study treatment(s) or receives another anticancer therapy prior to symptom deterioration. Only deaths occurring within 2 PRO assessment visits from the last available PRO assessment will be included as events.

Subjects whose symptoms (as measured by the QLQ-BIL21) have not shown a clinically meaningful deterioration and who are alive at the time of the analysis will be censored at the time of their last PRO assessment, where the symptom could be evaluated. Also, if symptoms progress or the subject dies after 2 or more missed PRO assessment visits, the subject will be censored at the time of the last PRO assessment, where the symptom could be evaluated prior to the 2 missed visits (refer to Table 13).

A sensitivity analysis of time to symptom deterioration will be conducted in which subject did not experience a clinically meaningful deterioration and deceased at the time of analysis will be censored at the last PRO assessment where the symptom could be evaluated or time of death if no PRO assessments where symptoms could be evaluated. Censoring will be applied in the same manner as for EORTC QLQ-C30 as described in Section 3.3.1.

The population for the analysis of time to symptom deterioration will include a subset of the PRO analysis set who have baseline scores ≤90.

Symptom improvement rate

Responses in symptoms for each visit (improvement, deterioration, and no change) based on Table 11 as well as the best overall response will be presented by treatment group. The symptom improvement rate will be defined as the number (%) of subjects with a best overall score response of "improved" in symptoms.

The denominator will consist of a subset of the PRO analysis set who have a baseline symptom score ≥10.

3.3.3 Patient reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE)

The patient reported outcomes version of the common criteria for adverse events (PRO-CTCAE), a PRO version of the CTCAE system developed by the National Cancer Institute (NCI), is included to assess tolerability from the subject's perspective. It was developed in recognition that collecting symptom data directly from subjects can improve the accuracy and efficiency of symptomatic AE data collection. Symptoms have been converted to subject terms (e.g., CTCAE term "myalgia" converted to "aching muscles"). Items capture the presence, frequency, severity and/or interference with usual activities, depending on the AE. Six items that are considered relevant for the trial were selected (CSP Appendix G). For each question, subjects select the value that best describes their experience over the past week.

PRO-CTCAE data will be presented using summaries and descriptive statistics based on the PRO analysis set. EORTC QLQ-BIL21 item 49 ("To what extent have you been troubled with side-effects from your treatment?") descriptive statistics will complement PRO-CTCAE findings.

3.3.4 Patient global impression of severity (PGIS)

The PGIS is a single item included to assess how a subject perceives their overall severity of symptoms at time of assessment. The response options of the PGIS are scored using a 6-point scale: 1 = No Symptoms; 2 = Very Mild; 3 = Mild; 4 = Moderate; 5 = Severe; 6 = Very Severe. PGIS data will be presented using summaries and descriptive statistics.

3.3.5 EQ-5D-5L

The EQ-5D-5L will be used to explore the impact of treatment and disease state on health state utility.

The EQ-5D-5L, developed by the EuroQol Group, is a generic questionnaire that provides a simple descriptive profile of health and a single index value for health status for economic appraisal. The EQ-5D-5L questionnaire comprises six questions that cover five dimensions of health (mobility, self-care, usual activities, pain/discomfort and anxiety/depression). For each dimension, respondents select which statement best describes their health on that day from a possible five options of increasing levels of severity (no problems, slight problems, moderate problems, severe problems and unable to/ extreme problems). A unique EQ-5D health state, termed the EQ-5D-5L profile, is reported as a five-digit code with a possible 3,125 health states. For example, state 11111 indicates no problems on any of the five dimensions. Respondents also assess their health today using the EQ-VAS, which ranges from 0 (worst imaginable health) to 100 (best imaginable health).

The EQ-5D profile will be converted into a weighted health state utility value, termed the EQ-5D index, by applying a country-specific equation to the EQ-5D-5L profile that represents the comparative value of health states. This equation is based on national valuation sets elicited from the general population and the base case will be the UK perspective. Where a valuation set has not been published, the EQ-5D-5L profile will be converted to the EQ-5D index using a crosswalk algorithm (Van Hout, 2012). The EQ-VAS is reported separately. Further details regarding the evaluation of EQ-5D-5L will be presented in the payer analyses plan (PAP). The evaluable population will comprise a subset of the PRO analysis set who have a baseline EQ-5D-5L assessment.

3.3.6 Compliance

Summary measures of compliance over time will be derived for all PRO questionnaires. These will be based upon:

- Received questionnaire: A questionnaire that has been received and has a completion date and at least one individual item completed.

- Expected questionnaire: A questionnaire that is expected to be completed at a scheduled assessment time e.g., a questionnaire from a subject who has not withdrawn from the study at the scheduled assessment time but excluding subjects in countries with no available translation.

    - For subjects that have progressed or discontinued study treatment, the earliest of date of study treatment discontinuation or progression will be used to determine the last on treatment windowed visit for each subjects expected forms using the analysis windows as described in Section 4.2.6.1. If the date falls before the end of the visit window, then that visit will only be considered expected if they have a received form. If they have not received a form, then this visit is not considered expected as they have not had the full opportunity to complete the questionnaire within the window. For subjects who have not discontinued study treatment or progressed, the date of the DCO will be used to determine the last on treatment visit for their last expected form following the same approach as above.

    - For follow up visits, if a subject has not discontinued study treatment then no follow up forms will be expected. For subjects who have discontinued study treatment, and discontinued the study, the date of study discontinuation will be used along with the visit windows for follow up day 30, follow up month 2 and follow up month 3 to determine the last expected visit that a form should have been completed. For subjects who have discontinued study treatment, and not discontinued the study, the date of the DCO will be used to determine whether follow up day 30, follow up month 2 and follow up month 3 are expected following the same approach as above.

- Expected forms (and compliance) will not be calculated beyond the 3 month follow up visit.
- Evaluable questionnaire: A questionnaire with a completion date and at least one subscale that is non-missing counted up until the earliest of progression and treatment discontinuation.

Compliance over time will be calculated separately for each visit up to month 3 follow-up, including baseline, as the number of subjects with an evaluable questionnaire at the time point (as defined above), divided by number of subjects still expected to complete questionnaires. Similarly, the evaluability rate over time will be calculated separately for each visit, including baseline, as the number of evaluable questionnaires (per definition above), divided by the number of received questionnaires.

3.4 Health care resource use variables

To investigate the impact of treatment and disease on health care resource of NON-STUDY protocol related events, the following variables will be captured in the HOSPAD form:

- Unplanned hospital attendances beyond trial protocol mandated visits (including physician visits, emergency room visits, day cases and admissions).
  • Type of attendance (outpatient/physician office attendance, hospitalization admission, and emergency room attendance)
- Primary sign or symptom the subject presents with.
- Length of hospital stay.
- Length of any time spent in an intensive care unit (ICU).

Where admitted overnight, the length of hospital stay will be calculated as the difference between the date of hospital discharge (or death date) and the start date of hospitalization or start of study drug if the start of study drug is after start date of hospitalization (length of hospital stay = end date of hospitalization – start date of hospitalization + 1). Subjects with missing discharge dates will be calculated as the difference between the last day with available data and the start date of hospitalization. The length of ICU stay will be calculated using the same method.

3.5 Safety variables

Safety and tolerability will be assessed in terms of adverse events (AEs) [including serious adverse events (SAEs)], deaths, physical examinations, laboratory findings, WHO/ECOG PS, vital signs, electrocardiograms (ECGs) and exposure, which will be collected for all subjects.

Data from all cycles of treatment will be combined in the presentation of safety. The SAF will be used for reporting of safety data, apart from deaths which are reported for FAS.

3.5.1 Study treatments

Study treatments/IP in this study are described in Table 15.

Table 16: Study treatments

	Durvalumab	**Placebo**	**Standard of care**
Study treatment name:	**Durvalumab (MEDI4736)**	**Sterile saline or dextrose solution**	**Standard of care (chemotherapy)[a]**
Dosage formulation:	500-mg vial solution for infusion after dilution, 50 mg/mL	Sterile solution of 0.9% (w/v) sodium chloride or 5% (w/v) dextrose for injection	As sourced locally
Route of administration:	IV	IV	IV
Dosing instructions[b]:	1500 mg IV q3w or q4w	0.9% (w/v) saline or 5% (w/v) dextrose volume matching durvalumab volume	Cisplatin 25 mg/m^2 and gemcitabine 1000 mg/m^2 on Day 1 and Day 8 q3w for up to 8 cycles
Packaging and labelling	Study treatment will be provided in 500-mg vials Each vial will be labelled in accordance with Good Manufacturing Practice (GMP)	Sourced locally by site	Sourced locally by site

	Annex 13 and per country regulatory requirement.[C]		
Provider	AstraZeneca	Sourced locally by site	Sourced locally by site[c]

[a] Under certain circumstances, when local sourcing is not feasible, an SoC (chemotherapy) treatment may be supplied centrally through AstraZeneca.

[b] Detailed instructions on IP administration are provided in Sections 6.1.1.1, 6.1.1.2, and 6.1.1.3 of CSP. Refer to Section 6.1.2 of CSP for details on the duration of treatment.

[c] Label text prepared for durvalumab (MEDI4736) will show the product name as "MEDI4736" or "durvalumab (MEDI4736)," depending upon the agreed product name used in the approved study master label document. All naming conventions are correct during this transitional period.

IV Intravenous; IP Investigational product; q3w Every 3 weeks; q4w Every 4 weeks; SoC Standard of care; w/v, weight/volume.

3.5.2 Exposure and dose interruptions

3.5.2.1 Treatment exposure for durvalumab or placebo

As durvalumab is initially dosed Q3W for up to 8 cycles in combination with gemcitabine/cisplatin (denoted as period 1 below), and then Q4W until clinical progression or RECIST 1.1-defined radiological PD (denoted as period 2 below), calculation of exposure (i.e. duration of treatment) will be defined as follows:

Total (or intended) exposure (months) of durvalumab or placebo: = (min(last durvalumab/placebo dose date where dose > 0 + [20 if last dose in period 1 or 27 if last dose in period 2], date of death, date of DCO) – first durvalumab/placebo dose date +1) /(365.25/12)

Actual exposure of durvalumab or placebo:

- Actual exposure = intended exposure – total duration of dose delays, where intended exposure will be calculated as above, and a dose delay is defined as any length of time where the subject has not taken any of the planned dose.

Dose interruptions – infusion

For durvalumab/placebo, a dose interruption is an infusion interruption that occurs during the infusion. The total dose received is >0. The drug can be restarted after the interruption and so it is possible for an infusion interruption to occur and the whole dose to still be administered. If the same infusion was interrupted multiple times, then this would just be captured as one infusion interruption.

Dose delays

A treatment cycle is started when >0 dose of durvalumab/placebo is administered. As such, a dose delay for durvalumab/placebo occurs when the start of a cycle is started at a later date than planned. If durvalumab/placebo is delayed, leading to other drugs that were scheduled to be administered on the same day being administered at a later date (but still the same day that durvalumab/placebo is eventually administered), then in this instance only durvalumab/placebo is classed as being delayed. This is because the other drugs would have been administered on the correct day relative to durvalumab/placebo. Dose reductions
Dose reductions are not permitted per CSP for durvalumab (or placebo). The actual exposure calculation makes no adjustment for any dose reductions that may have occurred.

Calculation of duration of dose delays (for actual exposure) will be defined as follows:

- Duration of durvalumab/placebo dose delay in period 1 or period 2 = Sum of (Date of the durvalumab/placebo dose in period 1 or period 2 - Date of previous durvalumab/placebo dose in respective period 1 or period 2 – [21days (if period 1) or 28 days (if period 2)]). If dose delay spans the transition between period 1 and period 2, duration of the durvalumab/placebo dose delay = Date of most recent durvalumab/placebo dose in period 2 – Date of last durvalumab/placebo dose in Period 1 – 28 days.

In the event that standard of care (SoC) is delayed due to SoC related toxicity, durvalumab may continue q3w up to 8 cycles followed by q4w. In that case the dosing schedule should be considered as period1 (q3W) for the calculation of durvalumab exposure.

In the event that SoC is permanently discontinued earlier than completion of 8 cycles due to SoC related toxicity, durvalumab may continue q4w. In that case dosing schedule should be considered as period2 (q4W) for the calculation of durvalumab exposure.

Number of treatment cycles received

Exposure will also be measured by the number of cycles received. A cycle corresponds to a period of 21 days (during period 1 whilst subject is receiving concomitant gemcitabine/ cisplatin) and 28 days (during period 2 which subject is receiving durvalumab or placebo alone). If a cycle is prolonged due to toxicity, this should still be counted as one cycle. A cycle will be counted if treatment is started even if the full dose is not delivered.

Subjects who permanently discontinue during a dose delay

If a subject permanently discontinues study treatment during a dose delay, then the date of last administration of study medication recorded on Exposure page will be used in the programming.

Safety Follow-up

Total Safety Follow-up = min ((last dose date +90), date of withdrawal of consent, date of death, date of DCO) – first dose date +1.

3.5.2.2 Treatment exposure for SoC (Gemcitabine plus Cisplatin)

Cisplatin and gemcitabine will be administered on Day 1 and Day 8 of each cycle (starting with cycle 1) for up to 8 q3w cycles. Exposure to gemcitabine and cisplatin will be defined as follows:

Total (or intended) exposure (months) of gemcitabine/cisplatin: = (min(last gemcitabine/ cisplatin dose date where dose > 0 + W, date of death, date of DCO) – first gemcitabine/ cisplatin dose date +1) /(365.25/12). Where W=6 if the last dose was scheduled on Day 1 and W=13 if the last dose was scheduled on Day 8.

Actual exposure of gemcitabine/cisplatin = intended gemcitabine/cisplatin exposure – total duration of gemcitabine/cisplatin dose interruptions, where intended exposure will be calculated as above, and a dose interruption is defined as any length of time where the subject has not taken any of the planned dose.

Duration of gemcitabine/ cisplatin dose delay = Sum of (Date of the gemcitabine/ cisplatin dose - Date of previous gemcitabine/cisplatin dose – X days). X=7 if Previous dose was on Day 1 of a cycle, X=14 if previous dose was on Day 8 of a cycle.

Dose modifications for gemcitabine/cisplatin should be followed local standard clinical practice. The number of dose delays/reductions/interruptions will be tabulated.

Dose reductions

Dose reductions, doses that are intentionally permanently reduced, are permitted as per CSP. This term is not used for interruptions or invalidly administered doses. A dose reduction is counted once for each time the dose is reduced.

Dose delays

A dose delay for gemcitabine or cisplatin occurs when the first administration of that drug (> 0 dose) in a cycle is administered at a later date relative to the durvalumab/placebo dose. Note that if the drug is completely skipped then this is not classed as a delay (it is classed as a dose interruption).

Dose interruptions - infusion

An infusion interruption of gemcitabine or cisplatin is defined the same as for durvalumab/placebo.

Dose interruptions – skipped doses

Since gemcitabine and cisplatin are administered multiple times per cycle, a skipped dose is a temporary interruption during a cycle. That is, during the cycle a dose is completely skipped or is taken at a later date than scheduled. Note that this is only applicable to drugs with multiple doses in a cycle (if the first dose is later than planned it would be a delay).

Number of treatment cycles (and Dose intensity) received in each arm should be analyzed.

The number of gemcitabine/cisplatin treatment cycles received will be calculated where a cycle corresponds to a period of 21 days.

3.5.3 Dose intensity

Dose intensity will be derived for study treatment including durvalumab, placebo, gemcitabine and cisplatin. Relative dose intensity (RDI) is the percentage of the actual dose intensity delivered relative to the intended dose intensity through to last day of dosing. RDI will be defined as follows:

- RDI = 100% * d/D, where d is the actual cumulative dose delivered up to the actual last day of dosing and D is the intended cumulative dose up to the actual last day of dosing. D is the total dose that would be delivered, if there were no modification to dose or schedule. When accounting for the calculation of intended cumulative dose 3 days should be added to the date of last dose to reflect the protocol allowed window for dosing.

When deriving actual dose administered the volume before and after infusion will also be used in the calculation.

Examples of dose intensity for durvalumab can be found in Table 17.

Table 17: Dose intensity scenarios for durvalumab

						Study Day					
RDI	Subject	1	22	43	64	85	106	127	148	169	
100%	1	X	X	X	X	X	X	X	X	X	PD
100%	2	X	X	X	X	X	X	X	X[D]		PD
55.6%	3	X		X		X	O		X	X	PD

						Study Day			
RDI	Subject	1	22	43	64	92	120	148	
77.8%	4	X	X	X	X	X	X	X	PD

X: Dose of 1500 mg taken; O: Dose Missed (delayed); [D]: Dose discontinued; PD: Progressive Disease

Subjects 1-4 progressed on Day 170, so the intended dose through to progression was 9×1500 mg of durvalumab = 13500mg.

Subject 1 received a total of 13500mg of durvalumab, whereas other subjects received less due to:

- Early stopping prior to PD (Subject 2)
- Dosing delays (Subject 3)
- SoC permanently discontinued earlier than completion of 8 cycles due to SoC related toxicity (Subject 4)

The Subject 2 example illustrates that for RDI, the end of actual dosing period is calculated based on the smallest recovery period after the last non-zero dose.

The Subject 4 example illustrates in the event that SoC is permanently discontinued earlier than completion of 8 cycles due to SoC related toxicity, however durvalumab is continuing per protocol q4w post discontinuation of SoC until PD (last dose).

Subject 1: RDI = (9x1500)/13500mg = 100%

Subject 2: RDI = (8x1500)/12000mg = 100%

Subject 3: RDI = (5x1500)/13500mg= 55.6%

Subject 4: RDI = (7x1500)/13500g = 77.8%

Placebo, gemcitabine and cisplatin RDI will be calculated in a similar manner to durvalumab according to the relevant dosing schedule of each treatment. The intended dose for durvalumab and placebo will be expected to be Q3W whilst the subject is receiving cisplatin followed by gemcitabine, after which the expected dose will be Q4W. The intended dose for gemcitabine and cisplatin will be Q3W with dosing on Day 1 and Day 8 of each cycle.

3.5.4 Adverse events (AEs)

An AE is the development of any untoward medical occurrence (other than progression of the malignancy under evaluation) in a subject or clinical study subject administered a medicinal product and which does not necessarily have a causal relationship with this treatment. An AE can therefore be any unfavorable and unintended sign (e.g., an abnormal laboratory finding), symptom (for example nausea, chest pain), or disease temporally associated with the use of a medicinal product, whether or not considered related to the medicinal product.

The term AE is used to include both serious and non-serious AEs and can include a deterioration of a pre-existing medical occurrence. An AE may occur at any time, including run-in or washout periods, even if no study treatment has been administered.

AEs and SAEs will be collected throughout this study. For this study, on treatment will be defined between date of start dose and 90 days following the last dose of study treatment or until the initiation of the first subsequent anti-cancer therapy (including radiotherapy, except for palliative radiotherapy) following discontinuation of study treatment (whichever occurs first). If an event starts outside of this period and it is considered possible that it is due to late onset toxicity to study drug, then it should be reported as an AE or SAE.

On treatment AEs (or treatment emergent AEs [TEAEs]) will be defined as any AEs that started after dosing or that started prior to dosing and worsened (by investigator report of a change in intensity) following exposure to treatment. If an AE is not worse than the baseline (pre-dose) severity, then it will not be classified at a TEAE.

The medical dictionary for regulatory activities (MedDRA) [using the latest or current MedDRA version] will be used to code AEs. AEs will be graded according to the National Cancer Institute (NCI) common terminology criteria for adverse event (CTCAE) version 5.0. The CTCAE grade will be assigned by the investigator as follows:

- Grade 1: Mild AE
- Grade 2: Moderate AE

- Grade 3: Severe AE
- Grade 4: Life-threatening or disabling AE
- Grade 5: Death related to AE

Missing start and stop dates for AEs will be handled using the rules described in Section 4.2.6.1. AEs that have missing causality (after data querying) will be assumed to be related to study.

3.5.5 Other significant adverse events (OAE)

During the evaluation of the AE data, an AstraZeneca medically qualified expert will review the list of AEs that were not reported as SAEs and 'Discontinuation of Investigational Product due to Adverse Events' (DAEs). Based on the expert's judgement, significant adverse events of particular clinical importance may, after consultation with the Global Subject Safety Physician, be considered other significant adverse events (OAEs) and reported as such in the CSR. A similar review of laboratory/vital signs/ECG data will be performed for identification of OAEs.

Examples of these are marked hematological and other laboratory abnormalities, and certain events that lead to intervention (other than those already classified as serious) or significant additional treatment.

3.5.6 AEs of special interest (AESI) and AEs of possible interest (AEPI)

Some clinical concepts (including some selected individual preferred terms) have been considered "AEs of special interest" (AESI) and "AEs of possible interest" (AEPI) to the durvalumab program. All AESIs are being closely monitored in clinical studies using durvalumab alone, and durvalumab in combination with other anti-cancer agents.

AESIs are defined as AEs that with a likely inflammatory or immune-mediated pathophysiological basis resulting from the mechanism of action of durvalumab and requiring more frequent monitoring and/or interventions such as corticosteroids, immunosuppressants, and/or endocrine therapy. Endocrine therapies include standard endocrine supplementation, as well as treatment of symptoms resulting from endocrine disorders (for example, therapies for hyperthyroidism include beta blockers [e.g., propranolol], calcium channel blockers [e.g., verapamil, diltiazem], methimazole, propylthiouracil, and sodium perchlorate). In addition, infusion-related reactions and hypersensitivity/anaphylactic reactions are also considered AESIs.

AEPIs are defined as AEs that could have a potential inflammatory or immune-mediated pathophysiological basis resulting from the mechanism of action of durvalumab but are more likely to have occurred due to other pathophysiological mechanisms, thus, the likelihood of the event being inflammatory or immune-mediated in nature is not high and/or is most often or usually explained by the other causes. These AEs not routinely arising from an inflammatory or immune-mediated mechanism of action – typically quite general clinical terms that usually present from a multitude of other causes –are classified as AEPIs.

These AESIs and AEPIs have been identified as Pneumonitis, Hepatic events, Diarrhea/Colitis, Intestinal perforations, Adrenal Insufficiency, Type 1 diabetes mellitus, Hyperthyroid events, Hypophysitis, Hypothyroid events, Thyroiditis, Renal events, Dermatitis/Rash (including pemphigoid), Pancreatic events, Myocarditis, Myasthenia gravis, Guillain-Barre syndrome, Myositis, Infusion/hypersensitivity reactions and Other rare/miscellaneous (including immune thrombocytopenia). Other categories may be added, or existing terms may be merged as necessary. An AstraZeneca medically qualified expert after consultation with the Global Patient Safety Physician has reviewed the AEs of interest and identified which MedDRA preferred terms contribute to each AESI/AEPI. A further review will take place prior to Database lock (DBL) to ensure any new terms not already included in the older MedDRA version are captured within the categories for the new higher MedDRA version. The list will be provided by AZ prior to database lock.

Immune-mediated adverse events

Durvalumab belongs to a class of drugs called immune checkpoint inhibitors. Because the mechanism of action of this class of drugs is to block the inhibitory signals that prevent T-cell activation, this drug may potentially cause immune-mediated adverse drug reactions (imAEs). An immune-mediated adverse event (imAE) is defined as an AESI that is associated with drug exposure and is consistent with an immune-mediated mechanism of action and where there is no clear alternate etiology. Infusion-related reactions and Hypersensitivity/Anaphylactic reaction AESIs are not considered for imAE adjudication because they are common to monoclonal antibody drugs in general and occur due to a mechanism of action different than that for imAEs.

imAE - automated adjudication

An automated adjudication process for defining whether AESI and AEPI are imAEs is based on applied rules and a treatment algorithm that considers interventions involving systemic steroid therapy, immunosuppressant use, and/or endocrine therapy (which, in the case of AEPIs, occurs after first considering an Investigator’s causality assessment and/or an Investigator’s designation of an event as immune-mediated). The algorithm referring to imAE treatment is given in the Durvalumab and Tremelimumab Global imAE Characterization Charter.

imAE - manual adjudication

A manual adjudication process may also be performed on AESI as outlined in the Durvalumab and Tremelimumab Immune-mediated Adverse Events (imAE) Characterization Charter.

3.5.7 Laboratory measurements

Laboratory data will be collected throughout the study as described in Tables 1 and 2 of the CSP. Blood and urine samples for determination of clinical chemistry, hematology, and urinalysis will be collected as described in Section 8.2.1 of the CSP.

For the derivation of baseline and post baseline visit values, the rules described in Section 4.1.1 of this document considering definition of baseline, visit windows and how to handle multiple records will be used.

Change from baseline in hematology and clinical chemistry variables will be calculated for each post-dose visit on treatment. CTCAE (version 5.0) grades will be defined at each visit according to the CTCAE grade criteria using local or project ranges as required, after conversion of lab result to corresponding AZ preferred units. The following parameters have CTCAE grades defined for both high and low values: Potassium, sodium, magnesium, glucose and corrected calcium so high and low CTCAE grades will be calculated.

Corrected calcium product will be derived during creation of the reporting database using the following formula:

Corrected calcium (mmol/L) = Total calcium (mmol/L) +([40 – albumin (G/L)] x 0.02)

Calculated creatinine clearance (CrCl) will be derived in the reporting database using the Cockcroft-Gault formula:

Creatinine clearance (mL/min) = ([140 – age at randomization] * weight (kg) [* 0.85 if subject is female]) / (72 * serum creatinine (mg/dL))

If weight is not available at a given visit in the reporting database weight from screening is used to derive calculated creatinine clearance. However, for summaries presented in TFLs calculated creatinine clearance will be rederived. If weight is not available at given visit then the weight from the previous available visit will be used in the calculation for TFLs summaries.

Absolute values will be compared to the project reference range and classified as low (below range), normal (within range or limits of range) and high (above range).

The maximum or minimum on treatment value (depending on the direction of an adverse effect) will be defined for each laboratory parameter as the maximum (or minimum) post-dose value at any time.

Project reference ranges will be used throughout for reporting purposes. If the project range is unavailable for a test, local ranges will be used. The denominator used in laboratory summaries of CTCAE grades will only include evaluable subjects (i.e., those who had sufficient data to have the possibility of an abnormality). For example,

- If a CTCAE criterion involves a change from baseline, evaluable subjects would have both a pre-dose and at least 1 post-dose value recorded.
- If a CTCAE criterion does not consider changes from baseline, to be evaluable the subject needs only to have 1 post dose-value recorded.

3.5.8 Vital signs

The following vital signs will be measured as described in Section 8.2.3 of the CSP: Systolic and diastolic blood pressure (BP), pulse rate, temperature, and respiratory rate. Body weight will also be recorded at each visit along with vital signs.

Vital signs will be collected at multiple times at same visit for the first infusion (pre-dose, during infusion, and at the end of infusion). At subsequent visits they may be taken at each of these timepoints as per institution and as clinically indicated.

Timepoints are reported by visit for each treatment arm, provided at least one treatment arm has ≥20 subjects with data at a given visit.

For the derivation of baseline and post-baseline visit values, the definitions and rules described in Section 4.2.6.1 for visit windows, and how to handle multiple records will be used.

Situations in which vital signs results should be reported as AEs are described in Section 8.3.7 of the CSP.

3.5.9 Physical examinations

Physical examinations will be performed as described in Section 8.2.2 of the CSP. Abnormalities recorded prior to the first dose of study treatment will be recorded as part of the subject’s baseline signs and symptoms. Abnormalities first recorded after first dose of study treatment will be recorded as AEs unless unequivocally related to the disease under study. Situations in which physical examination results should be reported as AEs are described in Section 8.3.7 of the CSP.

3.5.10 Electrocardiograms (ECGs)

Resting 12-lead ECGs will be recoded at screening and as clinically indicated throughout the study as described in Section 8.2.4 of the CSP.

The following ECG variables will be collected: ECG heart rate, PR duration, QRS duration, QT duration, RR duration and overall ECG evaluation.

The overall evaluation of an ECG will either be "normal" or "abnormal" with abnormalities categorized as either "clinically significant" or "not clinically significant". In case of clinically significant ECG abnormalities, 2 additional ECGs will be obtained over a brief period (e.g., 30 minutes) to confirm the finding.

The QT interval corrected for heart rate using Fridericia's correction (QTcF) will be calculated in the eCRF as follows (where QT and RR are in seconds):

$$\mathrm{QTcF} = \frac{\mathrm{QT}}{\sqrt[3]{\mathrm{RR}}}$$

Alternatively, RR (or QT) can be programmatically derived if not reported but QTcF and QT (or RR, respectively) is reported. RR can be calculated as follows:

$$\mathrm{RR} = \left(\frac{\mathrm{QT}}{\mathrm{QTcF}}\right)^3$$

Situations in which ECG results should be reported as AEs are described in Section 8.3.7 of the CSP.

3.5.11 World Health Organisation (WHO)/Eastern Cooperative Oncology Group (ECOG) performance status (PS)

The WHO/ECOG PS will be assessed as described in Section 8.2.5 of the CSP as the following:

0. Fully active; able to carry out all usual activities without restrictions

1. Restricted in strenuous activity, but ambulatory and able to carry out light work or work of a sedentary nature (e.g., light house work or office work)

2. Ambulatory and capable of self-care, but unable to carry out any activities; up and about more than 50% of waking hours
3. Capable of only limited self-care; confined to bed or chair more than 50% of waking hours
4. Completely disabled; unable to carry out any self-care and totally confined to bed or chair
5. Dead

Any significant changes from baseline or screening will be reported as AE.

3.6 Pharmacokinetic (PK) variables

PK concentration data will be collected as described in Section 8.5 of the CSP.

The actual sampling times will be used in the PK calculations. PK parameters, such as peak and trough concentration will be obtained from raw PK data measurements as data allow.

Individual concentrations below the Lower Limit of Quantification (LLOQ) of the bioanalytical assay will be reported as not quantifiable (NQ) in the listings with the LLOQ defined in the footnotes of the relevant TFLs. Individual serum concentrations that are Not Reportable will be reported as NR and those that are missing will be reported as NS (No Sample) in the listings. For data below limit of quantification (BLQ),NR or NS the following rules will apply:

- Any values reported as NR or NS will be excluded from the summary tables and corresponding figures.
- If, at a given time point, 50% or less of the serum concentrations are NQ, the geometric mean, CV%, geometric CV%, mean and SD will be calculated treating the NQ as LLOQ.
- If more than 50%, but not all, of the concentrations are NQ, the geometric mean, CV%, geometric CV%, and SD will be reported as data not calculable (NC). The maximum value will be reported from the individual data, and the minimum and median will be set to NQ.
- If all the concentrations are NQ, the geometric mean, mean, minimum, median and maximum will be reported as NQ and the CV%, geometric CV% and SD as NC.

3.7 Immunogenicity variables

Samples will be measured for the presence of ADAs (Anti-drug antibody) and neutralizing ADA (nAb) for durvalumab using validated assays. ADA sample analysis will be performed for both durvalumab and placebo treatment groups. Tiered analysis will be performed to include screening, confirmatory, titer and nAb assay components, and positive / negative cut points previously statistically determined from drug-naïve validation samples will be used. ADA data will be collected at scheduled visits as shown in the CSP (Section 8.5.2). ADA result from each sample will be reported as either positive or negative. If the sample is positive, the ADA titer will be reported as well. In addition, the presence of neutralizing ADA may be tested for all ADA-positive samples using a ligand-binding assay. The nAb results will be reported as positive or negative.

The number of subjects in the ADA analysis set who fulfil the following criteria will be determined. The percentage of subjects in each of the categories listed below will be calculated, using the number of subjects in the ADA analysis set of the treatment group as the denominator.

- ADA positive at any visit; the percentage of ADA-positive subjects in the ADA analysis set is known as ADA prevalence. A subject is defined as being ADA positive if a positive ADA result is available at any time, including baseline and all post-baseline measurements; otherwise ADA negative.

- Treatment-emergent ADA positive (either treatment-induced ADA positive or treatment-boosted ADA); the percentage of subjects fulfilling this criterion in the ADA analysis set is known as ADA incidence.

- ADA positive post-baseline and positive at baseline.

- ADA positive post-baseline and not detected at baseline (treatment-induced ADA positive).

- ADA not detected post-baseline and positive at baseline.

- Treatment-boosted ADA positive, defined as a baseline positive ADA titer that was boosted to a 4-fold or higher level (greater than the analytical variance of the assay) following drug administration.

- Treatment-emergent ADA persistently positive, defined as treatment-emergent ADA+ subjects having at least 2 post-baseline ADA positive measurements with at least 16 weeks (112 days) between the first and last positive measurement, or an ADA positive result at the last available assessment.

- Treatment-emergent ADA transiently positive, defined as treatment-emergent ADA+ subjects having at least one post-baseline ADA positive measurement and not fulfilling the conditions for TE-ADA persistently positive.

- nAb positive at any visit.

3.8 Biomarkers

Blood and tumor samples for exploratory biomarkers will be obtained according to the schedules presented in Section 1.1 of the CSP.

Pre-treatment tumor PD-L1 expression, as defined in the secondary objectives, will be evaluated (retrospectively) in all evaluable subjects. Data will be compared between arms to determine if baseline PD-L1 status is prognostic and/or predictive of outcomes associated with Arm A versus Arm B. Other exploratory biomarkers, such as tissue and/or blood based tumor mutational burden and microsatellite instability (MSI)/mismatch repair proficiency will also be evaluated retrospectively. Detailed description of biomarker data can be found in Section 8.8 of CSP.

3.9 Other variables

3.9.1 Prior and concomitant medications and therapies

All therapies (drug and non-drug), including herbal preparations, whether prescribed or over-the-counter, that are used within the four weeks prior to initiation of study treatment up until 90 days following last dose of study treatment will be recorded on the eCRF. Details include generic and/or brand names of the medications, World Health Organization Drug Dictionary (WHO-DD) encoding (using the latest or current WHO-DD version), reason for use, route, dose, dosing frequency, and start and stop times.

Prior medications are those taken prior to study treatment.

Concomitant medications are those with a stop date on or after the first dose date of study treatment or ongoing (and could have started prior to or during treatment).

Missing start and stop dates for medications will be handled using the rules in Section 4.2.6.1. Missing coding terms should be listed and summarized as "Not coded".

4 ANALYSIS METHODS

The primary objective of the study is to confirm the superiority of durvalumab plus gemcitabine/cisplatin combination therapy (Arm A) compared to placebo plus gemcitabine/cisplatin therapy (Arm B) in terms of OS in subjects with previously untreated, unresectable locally advanced or metastatic BTC.

Results of all statistical analysis will be presented using a 95% confidence interval (CI) and 2-sided p-value, unless otherwise stated.

The formal statistical analysis will be performed to test the following main hypotheses:

- H0: No difference between Arm A and Arm B
- H1: Difference between Arm A and Arm B

There will be 3 data cut-offs (DCO) for this study consisting of 2 interim analyses and 1 final analysis. This study will have met its primary objective if Arm A is statistically significantly superior to Arm B, either at IA-2 or at the final analysis.

1. **Interim Analysis-1 (IA-1)**: The objective of IA-1 is to assess clinical activity. ORR and DoR will be summarized to support early registration of durvalumab when administered in combination with gemcitabine/cisplatin. The summaries will be done both for Investigator assessments and for blinded independent central review (BICR) assessments according to RECIST 1.1. The BICR summaries will be of primary interest with the Investigator data providing supportive evidence. The planned DCO for IA-1 will occur when at least 200 subjects have completed at least 32 weeks of follow-up or the last subject has been randomized to the global cohort whichever comes later. The analysis set will include all randomized subjects who have had the opportunity for at least 32 weeks of follow-up at the time of the IA-1 DCO (FAS-32w, i.e. randomized ≥32 weeks prior to IA-1 DCO).
   Based on enrolment assumptions, it is expected that this will occur approximately 21months after randomization of the first subject.

2. **Interim Analysis -2 (IA-2):** IA-2 will test for early superiority of the durvalumab regimen relative to control. This analysis will be performed when approximately 397 OS events have been observed in the study (59% maturity or 80% information fraction). Based on enrolment assumptions, it is expected that this will occur approximately 31 months after randomization of the first subject.

3. **Final Analysis (FA):** The FA will be performed when approximately 496 OS events have been observed in the study (74% maturity). Based on enrolment assumptions it is anticipated that this analysis will be performed 40 months after the first subject is randomized.

Refer to Section 5 for further details of planned interim analysis.

4.1 General principles

Efficacy data will be summarized and analyzed on the FAS. PRO data will be analyzed on the PRO analysis set. Safety and treatment exposure data will be summarized based upon the SAF. Study population and demography data will be summarized based upon the FAS. PK data will be analyzed using the PK analysis set. Study day for efficacy analyses will be relative to the randomization date. Study day for safety analyses and PROs will be relative to the date of first dose of study treatment. For subjects randomized and not treated, randomization date will be used instead to assign study day for PRO endpoints.

The below mentioned general principles will be followed throughout the study:

- All analyses and reporting will be by treatment arm.
- Descriptive statistics will be used for all variables, as appropriate. Continuous variables will be summarized by the number of observations, mean, standard deviation, median, upper and lower quartiles minimum, and maximum. For log-transformed data it is more appropriate to present geometric mean, coefficient of variation (CV), median, minimum and maximum. Categorical variables will be summarized by frequency counts and percentages for each category.
- Unless otherwise stated, percentages will be calculated out of the population total for the corresponding treatment group. Overall totals will be calculated for baseline summaries only.
- For continuous data, the mean and median will be rounded to 1 additional decimal place compared to the original data. The standard deviation will be rounded to 2 additional decimal places compared to the original data. Minimum and maximum will be displayed with the same accuracy as the original data.
- For categorical data, percentages will be rounded to 1 decimal place.
- In general, unless otherwise stated for subgroups, analysis will not be performed if there are < 5 subjects in a subgroup. Descriptive summaries may still be provided. For the number of events required for meaningful analysis of subgroups for OS and PFS refer to section 4.2.2
- SAS® version 9.1 (or higher) will be used for all analysis.

4.1.1 Definition of baseline

In general, for efficacy endpoints the last observed measurement prior to randomization will be considered the baseline measurement. However, if an evaluable assessment is only available after randomization but before the first dose of randomized treatment then this assessment will be used as baseline. For safety and PRO endpoints, the last observation before the first dose of study treatment will be considered the baseline measurement unless otherwise specified. For subjects randomized and not treated, randomization date will be used instead to assign baseline measurement and study day for PRO endpoints. For assessments on the day of first dose where time is not captured, a nominal pre-dose indicator, if available, will serve as sufficient evidence that the assessment occurred prior to first dose.

Assessments on the day of the first dose where neither time nor a nominal pre-dose indicator are captured will be considered prior to the first dose if such procedures are required by the protocol to be conducted before the first dose.

For safety endpoints baseline will be defined as the last non-missing measurement of the variable under consideration prior to the intake of the first dose of study treatment. That is, the latest result prior to the start of study treatment. If two visits are equally eligible to assess subject status at baseline (e.g., screening and baseline assessments both on the same date prior to the first dose with no washout or other intervention in the screening period), the average will be used as the baseline value. For non-numeric laboratory tests (i.e., some of the urinalysis parameters) where taking the average is not possible, the best value would be taken as baseline as this is most conservative (the order from the best to the worst is: NEG, TRACE, POS, 0, +, ++, +++, >+++). In the scenario where there are two assessment recorded on the day, one with time recorded and the other without time recorded, the one with the time recorded would be selected as baseline. Where safety data are summarized over time, time on study will be calculated in relation to date of first study treatment.

In all summaries change from baseline variables will be calculated as the post-treatment value minus the value at baseline. The percentage change from baseline will be calculated as (post-baseline value - baseline value) / baseline value x 100.

Unless otherwise specified, date of initiation of the first subsequent therapy should be the date of the first subsequent anti-cancer therapy (excluding radiotherapy). Assessments on the day of first subsequent therapy will be considered prior to start of subsequent therapy.

4.2 Analysis methods

Table 18 details which endpoints are to be subjected to formal analysis, together with pre-planned sensitivity analyses, making it clear which analysis is regarded as primary for that endpoint.

Table 18: Formal statistical analyses to be conducted and pre-planned sensitivity analyses

Endpoints analyzed	Notes
Overall survival	Primary confirmatory analysis IA-2: Stratified log-rank analysis test adjusting for disease status and primary tumor location for primary comparison of survival between randomized treatment groups providing a p-value and stratified Cox proportional hazard model providing hazard ratio (HR) (95% CI) and ([1-adjusted alpha] x 100%) FA: Stratified FH(0, 1) test adjusting for disease status and primary tumor location for primary comparison of survival between randomized treatment groups providing a p-value and stratified Cox proportional hazard model providing hazard ratio (HR) (95% CI) and ([1-adjusted alpha] x 100%) Sensitivity and supplemental analysis KM plot of time to censoring where the censoring indicator of the primary analysis is reversed – attrition bias Cox proportional hazards models to determine the effect of covariates on the HR estimates Subgroup analysis using Cox model Sensitivity analysis at FA: Stratified log-rank test adjusting for disease status and primary tumor location for primary comparison of survival between randomized treatment groups

Endpoints analyzed	Notes
Progression free survival	Only PFS according to RECIST 1.1 based on investigator assessments will be analyzed as a secondary variable in a confirmatory manner Secondary confirmatory analysis Stratified log-rank tests adjusting for disease status and primary tumor location, using PFS according to RECIST 1.1 using Investigator assessments providing a p-value and stratified Cox proportional hazard model providing hazard ratio (HR) (95% CI) Sensitivity and supplemental analysis Interval-censored analysis – evaluation time bias Analysis using alternative censoring rules – attrition bias Cox proportional hazard models to determine the effect of covariates on the HR estimate Subgroup analysis using Cox proportional hazard model
Objective response rate	IA-1: Exact Clopper-Pearson confidence intervals and a p-value from a stratified CMH test adjusting for disease status and primary tumor location Primary analysis with tumor data according to RECIST 1.1 based on BICR in FAS-32w with a measurable disease at baseline per BICR. Sensitivity analysis in a subset of FAS-32w IA-2 and FA: Odds ratio and p-value from a CMH test adjusted for disease status and primary tumor location, using tumor data according to RECIST 1.1 by Investigator assessment
Duration of response	KM plot and Swimmer plot of DoR according to RECIST 1.1 based on Investigator assessments. Median DoR calculated from the KM curve. At IA-1, KM plot and Swimmer plot of DoR according to RECIST 1.1 as assessed by BICR

Endpoints analyzed	Notes
Disease control rate	Summary statistics using DCR, DCR-24w, DCR-32w and DCR-48w as assessed by the Investigator according to RECIST 1.1 At IA-1, summary statistics using DCR, DCR-24w, DCR-32w and DCR-48w as assessed by BICR
Summary and descriptive statistics for each scale/item: EORTC QLQ-C30 and QLQ-BIL21	Summary and descriptive statistics Unadjusted change from baseline
Change from baseline in symptoms, functions, global health status/QoL domains or items: EORTC QLQ-C30 and QLQ-BIL21	Adjusted mean change from baseline using MMRM analysis (overall and by each visit)
Time to symptom, function, or global health status/QoL deterioration: EORTC QLQ-C30 and QLQ-BIL21	Stratified log-rank test (for p-value), HR from Cox model (with 95% CI), KM plot
PRO improvement rates for symptoms, functions or global health status/QoL domains or items: EORTC QLQ-C30 and QLQ-BIL21	Logistic regression with odds ratio, 95% CI and p-value
Patients global and treatment-related symptoms: PGIS, PRO-CTCAE. QLQ-BIL21 item	Summary descriptive statistics
EQ-5D-5L (health state utility values and Visual Analog Scale)	Summary statistics for health state utilities and visual analogue scale, including change from baseline.

Endpoints analyzed	Notes
Healthcare resource use	Descriptive statistics (as appropriate, including means, median, ranges or frequencies and percentages)

BICR Blinded independent central review; CR Complete response; DCR Disease control rate; DCR-24w Percentage of subjects who have a best objective response of CR or PR or who have SD for at least 24 weeks (±7 days), following the start of study treatment; DCR-32w Percentage of subjects who have a best objective response of CR or PR or who have SD for at least 32 weeks (±7 days), following the start of treatment; DoR Duration of response; EORTC European Organisation for Research and Treatment of Cancer; FAS Full analysis set; HR Hazard ratio; HRQoL Health related quality of life; KM Kaplan Meier; MMRM Mixed-effect model repeated measure; OS Overall survival; PFS Progression free survival; PGIS Patient Global Impression of Severity; PR Partial response; PRO Patient reported outcomes; PRO-CTCAE Patient reported outcomes Common Terminology Criteria for Adverse Events; QLQ-BIL21 21-Item Cholangiocarcinoma and Gallbladder Cancer Quality of Life Questionnaire; QLQ-C30 30-Item Core Quality of Life Questionnaire; QoL Quality of life; RECIST Response Evaluation Criteria in Solid Tumors; SD Stable disease .

4.2.1 Multiplicity

A small alpha expenditure of 0.001 (0.1%) will be allocated to IA-1 for ORR. Strong control of the FWER at the remaining 4.9% level (2-sided) across the testing of OS and PFS endpoints will be achieved through a combined approach of alpha allocation to the OS analyses (IA-2 and the FA) via O’Brien Fleming alpha spending function and a hierarchical testing procedure; that is, PFS will be tested only if OS met statistical significance at IA-2 or FA (Glimm et al. 2010). The IA-2 for OS will be conducted when approximately 397 of the 496 expected OS events (i.e., 80% information fraction) have occurred. The significance level for the primary OS analysis at IA-2 will be decided using the Lan-DeMets spending function approximating O’Brien-Fleming boundaries (Lan and DeMets 1983). If approximately 397 of 496 expected OS events are observed at IA-2, 2-sided significance levels of 0.0238 will be applied to the primary OS analysis at IA-2 using log-rank test.

The statistical significance for the primary OS analysis at FA using FH(0, 1) will be determined based on the alpha spending at IA-2 and the correlation structure between IA-2 log-rank test statistic and FA FH(0, 1) test statistic based on the actual data collected at FA(Tsiatis 1982). Let S_1, S_2 be the score statistics of the logrank test at IA2 and FH(0, 1) test at FA, then S_1 and S_2 have the following covariance matrix.

$$\hat{\sigma}_1^2 = \int_0^{t_1} \hat{Q}_1^2(t_1, u) \frac{Y_1(u)Y_0(u)}{Y_1(u) + Y_0(u)} \left(1 - \frac{dN_1(u) + dN_0(u) - 1}{Y_1(u) + Y_0(u) - 1}\right) \frac{dN_1(u) + dN_0(u)}{Y_1(u) + Y_0(u)}$$

$$\hat{\sigma}_2^2 = \int_0^{t_2} \hat{Q}_2^2(t_2, u) \frac{Y_1(u)Y_0(u)}{Y_1(u) + Y_0(u)} \left(1 - \frac{dN_1(u) + dN_0(u) - 1}{Y_1(u) + Y_0(u) - 1}\right) \frac{dN_1(u) + dN_0(u)}{Y_1(u) + Y_0(u)}$$

$$\hat{\sigma}_{12} = \int_0^{t_1} \hat{Q}_1(t_1,u)\hat{Q}_2(t_2,u)\frac{Y_1(u)Y_0(u)}{Y_1(u)+Y_0(u)}\left(1-\frac{dN_1(u)+dN_0(u)-1}{Y_1(u)+Y_0(u)-1}\right)\frac{dN_1(u)+dN_0(u)}{Y_1(u)+Y_0(u)},$$

Where t_1 is the analysis data cutoff (DCO) time for IA2 and t_2 is the DCO at final analysis, and t_1 and t_2 are calculated from the first subject randomization date. In addition, $\hat{Q}_1(t_1,u) = 1$ for the standard logrank test, $\hat{Q}_2(t_2,u)$= 1- $\hat{S}(t_2,-u)$ for FH(0, 1) test at FA, where $\hat{S}(t_2,-u)$ is the pooled KM estimate of the survival rate just before survival time u. $Y_1(u)$ and $Y_0(u)$ denote the number of subjects at risk just before time u in the experimental arm and control arm, respectively. $N_1(u)$ and $N_0(u)$ denote the counting process of event(s) at survival time u in the experimental arm and control arm, respectively. The correlation between the normalized logrank test statistic at IA2 and FH(0, 1) test statistic at FA can be determined accordingly, $\rho = \hat{\sigma}_{12}\left(\hat{\sigma}_2^2\hat{\sigma}_2^2\right)^{-1/2}$.

The log-rank score statistic at IA2 and FH(0, 1) score statistic at final analysis follow asymptotical bivariate normal distribution with the above covariance structure. As a result, the rejection boundary for FH(0, 1) test at final analysis can be determined according to the alpha spending at interim and this covariance structure (Tsiatis 1982, Prior 2020). More in-depth discussions of the method are available at (He et al. 2021). The overall alpha can be strongly controlled using the group sequential test method based on the correlation:

$$P(|Z_{LR}| > z_1|H_0) = \alpha_1 \text{ and } P(|Z_{LR}| < z_1, |Z_{FH01}| > z_F|H_0) = \alpha - \alpha_1$$

where α_1 is the allocated type I error for IA2 for the logrank test, and $\alpha - \alpha_1$ is the allocated type I error for FA using FH(0, 1) when H_0 is not rejected at IA2 using logrank test; and z_1 and z_F are the rejection bounds at IA2 and FA respectively.

PFS will be formally tested using PFS information collected up to each DCO if OS meets statistical significance at that DCO (IA-2 or FA). Significance levels for PFS at IA-2 and FA for the log-rank test will be derived based on the Lan-DeMets alpha spending function approximating Pocock boundaries, which strongly controls the Type I error at the 0.049 level (2-sided). Assuming approximately 506 PFS events and 590 PFS events are available at the time of each PFS analysis, PFS testing will be carried out with 2-sided significance levels of 0.0444 and 0.0236 at IA-2 and FA for the log-rank test, respectively. Since DCO timing will be determined based on the number of OS events, the nominal significance level for PFS analysis might be adjusted for the actual information fraction for PFS at IA-2 relative to FA. The significance levels for the log-rank test will be calculated using EAST for OS and PFS at the time of the interim and final analyses.

Simulation studies were performed with 100,000 runs for each scenario in Table 19 and Table 20, and the precision level (1.96*se) of the overall type I error is $1.96\left(\frac{0.0245(1-0.0245)}{100000}\right)^{1/2} \approx$

0.001. The simulations demonstrate strong type I error control using the proposed method, logrank for IA and FH(0, 1) for FA, regardless of the following considerations: (1) Accrual patterns (scenarios A1-A4); (2) Distributions of control arm (scenarios D1-D6); (3) Timing of analyses for IA and FA (scenarios E1-E4); (4) Sample size (scenarios S1-S2).

The results are summarized in Table 19 and Table 20 below. The random samples for (piecewise) exponential distributions are generated using R package *nphsim*. The statistical inference is based on an internal R package *wlr*(He et al. 2021).

The remaining incremental alpha from IA-2 will be used for the final analysis, and the alpha boundary will be able to be identified at that time. This is in accordance with the Tsiatis 1982 publication in that the correlation between the log-rank test at IA-2 and final analysis will be determined based on actual pooled data at the time of final analysis.

Table 19: Type I error (1-sided) for IA and overall study by simulations for various accrual patterns and distributions

Scenarios	Distribution	Accrual Pattern [a]	IA: Logrank FA: Logrank		IA: Logrank FA: $FH(0,1)$	
			IA	Overall	IA	Overall
A1	$\exp(\lambda = \frac{\log(2)}{11.7})$	21 mo with $\xi = 1.5$	0.01166	0.02420	0.01166	0.02436
A2	same as above	21 mo with $\xi = 2$	0.01139	0.02457	0.01139	0.02427
A3	same as above	24 mo with $\xi = 1.5$	0.01184	0.02484	0.01184	0.02407
A4	same as above	24 mo with $\xi = 2$	0.01141	0.02424	0.01141	0.02416
D1	$\exp(\lambda = \frac{\log(2)}{15})$	21 mo with $\xi = 1.5$	0.01221	0.02466	0.01221	0.02528
D2	Piecewise exp. 1 [b]	same as above	0.01230	0.02507	0.01230	0.02553
D3	Piecewise exp. 2 [c]	same as above	0.01261	0.02562	0.01261	0.02499
S1	$\exp(\lambda = \frac{\log(2)}{11.7})$	same as above $n = 620$	0.01169	0.02478	0.01169	0.02472
S2	same as above	same as above $n = 720$	0.01181	0.02414	0.01181	0.02454

Note: The simulation results for each scenario are based on 100,000 runs. Each run has a total sample size of 672 subjects with randomization 1:1 for scenarios A1-A4 and D1-D3. The overall type I error (1-sided) is 0.0245. Based on O'Brien Fleming spending function, the type I error (1-sided) 0.01194 is allocated to IA for planned IA and FA being performed at 397 and 496 events respectively, per study design. For scenarios S1-S2, the target events for IA and FA are 397 and 496 respectively.

[a] Accrual pattern: the proportion of cumulative enrollment at time t is $\left(\frac{t}{A}\right)^{\xi}$ for $0 < t \leq A$. For uniform enrollment, $\xi = 1$. Larger ξ means more accelerated enrollment rate at later time.

[b] Piecewise exponential distribution with hazards $\lambda_1 = \frac{0.7\log(2)}{11.7}$ before 6 months, and $\lambda_2 = \frac{\log(2)}{11.7}$ after 6 months.

[c] Piecewise exponential distribution with hazards $\lambda_1 = \frac{\log(2)}{11.7}$ before 6 months, and $\lambda_2 = \frac{0.7\log(2)}{11.7}$ after 6 months.

Table 20: Type I error (1-sided) for IA and overall study by simulations when events are deviated from planned target events

Scenario	Timing of Analysis		IA: Logrank FA: Logrank		IA: Logrank FA: FH(0, 1)	
			IA	Overall	IA	Overall
E1	360, 496	Allocated type I error [a]	0.00829	0.02450	0.00829	0.02450
		Simulation	0.00837	0.02507	0.00837	0.02550
E2	420, 496	Allocated type I error [a]	0.01452	0.02450	0.01452	0.02450
		Simulation	0.01500	0.02460	0.01500	0.02484
E3	360, 530	Allocated type I error [a]	0.00635	0.02450	0.00635	0.02450
		Simulation	0.00629	0.02440	0.00629	0.02411
E4	420, 530	Allocated type I error [a]	0.01152	0.02450	0.01152	0.02450
		Simulation	0.01207	0.02468	0.01207	0.02534
D4	360, 496	Piecewise exp. 1 [b]	0.00808	0.02498	0.00808	0.02458
D5	360, 496	Piecewise exp. 2 [c]	0.00827	0.02406	0.00827	0.02436
D6	360, 496	$\exp(\lambda = \frac{\log(2)}{15})$	0.00797	0.02369	0.00797	0.02412

Note: Exponential distribution with median 11.7 months and accrual pattern is 21 months with weight 1.5. [a] Allocated type I error according to O'Brien Fleming spending function. For scenarios D4-6, the allocated type I error is the same as scenario E1.

4.2.2 Primary efficacy endpoint overall survival (OS)

The primary endpoint OS will be analyzed using a stratified log-rank at IA-2 and using a stratified FH(0, 1) test at FA for the generation of p-value. Both analyses will adjust for disease status (initially unresectable or recurrent) and primary tumor location (intrahepatic cholangiocarcinoma, extrahepatic cholangiocarcinoma, or gallbladder cancer). FH(0, 1) test can be programmed in SAS using "TEST=FH(0, 1)" option in STRATA statement of PROC LIFETEST. To estimate the effect of treatment, the HR together with its 95% CI and ([1-adjusted alpha] x 100%) will be estimated from a stratified Cox proportional hazards model (Cox, 1972) with ties = Efron and the stratification variables included in the strata statement and the CI calculated using the profile likelihood approach. OS at month 12, month 18 and month 24 will also be summarized (using Kaplan-Meier curve) and presented by treatment arm.

The stratification variables in the statistical modelling will be based on the values entered into IVRS/IWRS at randomization, even if it is subsequently discovered that these values were incorrect.

Kaplan-Meier (KM) plots of OS will be presented by treatment arm. Summaries of the number and percentage of subjects who have died, those still in survival follow-up, those lost to follow-up and those who have withdrawn consent will be provided along with the median OS for each treatment.

Assumptions of proportionality

The assumption of proportionality will be assessed. Proportional hazards will be tested firstly by examining plots of complementary log-log (event times) versus log (time) and, if these raise concerns, by fitting a time dependent covariate (adding a treatment-by-time or treatment-by-ln(time) interaction term) to assess the extent to which this represents random variation. If a lack of proportionality is evident, the variation in treatment effect will be described by presenting piecewise HR calculated over distinct time-periods. In such circumstances, the HR from the primary analysis can still be meaningfully interpreted as an average HR over time unless there is extensive crossing of the survival curves. If lack of proportionality is found this may be a result of a treatment-by-covariate interaction, which will be investigated. In addition, the KM curve along with landmark analyses (e.g., 1-year OS rate) will also help in understanding the treatment benefit.

Sensitivity and supplemental analysis

The following sensitivity and supplemental analysis will be performed.

Attrition bias

A sensitivity analysis for OS will examine the censoring patterns to rule out attrition bias with regard to the primary treatment comparisons, achieved by a Kaplan-Meier plot of time to censoring where the censoring indicator of OS is reversed.

The number of subjects prematurely censored will be summarized by treatment arm. A subject would be defined as prematurely censored if their last known alive date is prior to DCO.

In addition, duration of follow-up will be summarized using medians:

- In censored subjects who are alive at DCO only: Time from randomization to date of censoring (date last known to be alive) for each arm.

Effect of covariates on the HR estimate (Cox proportional hazards model)

Cox proportional hazards modelling will be used to assess the effect of covariates on the HR estimate. A model will be constructed, containing treatment, the stratification factors and the following covariates age, sex, race, ECOG and locally advanced versus metastatic BTC. Where ECOG status is from screening.

Interactions between treatment and stratification factors will also be tested to rule out any qualitative interaction using the approach of Gail and Simon 1985.

At FA OS will also be analyzed using a stratified log rank test, adjusting for disease status (initially unresectable or recurrent) and primary tumor location (intrahepatic cholangiocarcinoma, extrahepatic cholangiocarcinoma, or gallbladder cancer) as sensitivity analysis.

Subgroup analysis

Subgroup analyses will be conducted comparing OS between of durvalumab plus gemcitabine/cisplatin combination therapy (Arm A) versus of placebo plus gemcitabine/cisplatin combination therapy (Arm B) in the following subgroups of the FAS (but not limited to):

- Sex (male versus female)
- Age at randomization (<65 versus ≥65 years of age)
- PD-L1 expression (see Section 4.2.9 for the subgroup definition)
- Disease status (initially unresectable versus recurrent) based on the values entered into eCRF at randomization
- Primary tumor location (intrahepatic cholangiocarcinoma versus extrahepatic cholangiocarcinoma versus gallbladder cancer) based on the values entered into eCRF at randomization
- Race (Asian versus non-Asian)
- Region (Asia versus Rest of the World)
- WHO/ECOG PS 0 versus 1 at screening
- Locally advanced versus metastatic BTC
- MSI status (Microsatellite instability-high versus Microsatellite stable) if there is adequate sample size in MSI high group

For these subgroup analyses any subject with missing values will be excluded from that particular subgroup.

Other baseline variables may also be assessed if there is clinical justification or an imbalance is observed between the treatment arms. The purpose of the subgroup analyses is to assess the consistency of treatment effect across expected prognostic and/or predictive factors. Forest plots will be presented.

No adjustment to the significance level for testing of the subgroup and sensitivity analyses will be made, since all these analyses will be considered supportive of the analysis of OS and PFS. For each subgroup level of a factor, the HR (for the treatment comparisons of interest) and 95% CI will be calculated from a Cox proportional hazards model that only contains a term for treatment. The Cox models will be fit using a SAS PROC PHREG with the Efron method to control for ties, using the by statement to obtain HR and 95% CI for each subgroup level separately. These will be presented on a forest plot including the HR and 95% profile likelihood CI, along with the results of the overall primary analysis.

If there are too few events available for a meaningful analysis of a particular subgroup comparison (it is not considered appropriate to present analyses where there are less than 20 events within a subgroup category (i.e., when the events in the treatment comparison does not add up to 20) in a subgroup), the relationship between that subgroup and the primary endpoint (OS) will not be formally analyzed. In this case, only descriptive summaries will be provided.

Consistency of treatment effect between subgroups

The presence of quantitative interactions between treatment and stratification factors will be assessed by means of an overall global interaction test for plausible subgroups.

This is performed by comparing the fit of a Cox proportional hazards model including treatment, all covariates, and all covariate-by treatment interaction terms, with one that excludes the interaction terms, and will be assessed at the 2-sided 10% significance level. If there are not more than 10 events per stratum for any covariate (i.e., within each stratum of a treatment*covariate interaction [2 treatments * 2 levels of the covariate = 4 stratum]) a pre-defined pooling strategy should be applied to the covariate. If the pooling strategy does not meet the event criteria, then the covariate-by-treatment interaction term should be omitted from the model. Moreover, if the covariate does not have more than 10 events per level of covariate then the main effect of the covariate will also be excluded. If the fit of the model is not significantly improved, then it will be concluded that overall the treatment effect is consistent across the subgroups.

If the global interaction test is found to be statistically significant, an attempt to determine the cause and type of interaction will be made. Stepwise backwards selection will be performed on the saturated model, whereby (using a 10% level throughout) the least significant interaction terms are removed one-by-one and any newly significant interactions re-included until a final model is reached where all included interactions are significant, and all excluded interactions are non-significant. Throughout this process all main effects will be included in the model regardless of whether the corresponding interaction term is still present. This approach will identify the factors that independently alter the treatment effect and prevent identification of multiple correlated interactions.

Any quantitative interactions identified using this procedure will then be tested to rule out any qualitative interaction using the approach of Gail and Simon 1985.

The mechanism of action of immunotherapy is to harness the immune system to recognize and destroy tumor cells instead of directly killing them with chemotherapy or radiation (Barrueto et al 2020). This indirect mechanism requires the time to mount an effective immune response, and the time for that response to be translated into an observable clinical response (Hoos 2012). Thus, delayed separation of survival curves may be observed between the experimental and control groups in clinical trials with time-to-event endpoints violating the proportional hazards assumption (Xu et al 2018). To date, out of the 20 randomized controlled Phase 3 studies published, a delay in survival curve separation was observed in 80% of these trials with 13 showing a delayed treatment effect of at least 3 months.

The commonly used statistical methods to analyze time to event endpoints such as overall survival and progression-free survival are the log-rank test for statistical inference and Cox proportional hazards model for quantification of treatment effect under the assumption of proportional hazards.

From a statistical perspective, this situation results in the violation of the proportional hazards assumption. The standard log-rank test, although optimal under proportional hazards (Schoenfeld 1981), suffers substantial power loss in handling survival data with delayed treatment effect (Lin et al. 2020). The use of weighted log-rank tests accounting for delayed treatment effects has been the subject of an increasing number of publications in recent years.

This feature is also recognized in regulatory guidance documents including FDA Guidance for Industry: Clinical considerations for therapeutic cancer vaccines, 2011 (FDA 2011) and two guidance documents provided on the PMDA website: Peptide Vaccine Guidance for Cancer Treatment (Yamaguchi et al 2014) and Guidance for developing cancer immunotherapy in late phase clinical trials (The Review Committee for Guidance Development 2018). Therefore, the use of a weighted log-rank test like the FH class has been proposed as an alternative analysis method for survival endpoints.

For cases where a delay in separation is expected, the method of statistical testing utilizing a weighted log-rank method of FH(0, 1) can better capture the statistical inference of whether durvalumab + Gemcitabine/Cisplatin is superior to placebo + Gemcitabine/Cisplatin in the TOPAZ-1 study. One could have extensive follow-up in order to boost the power of log-rank test, but further exploration indicates that at least 12 months of additional follow-up are needed in order to achieve the equivalent power provided by FH(0, 1) test and the maturity would be as high as 88%, which would substantially result in delay in the availability of a novel treatment option for a subject population with high unmet need.

Fleming-Harrington (FH) is one class of weighted log-rank test (Fleming and Harrington 1991), FH(ρ, γ), which assigns weight according to the survival rate $\hat{S}(u)$ estimated based on pooled data from two treatment arms at survival time u.

$$\hat{Q}(u) = \left(\hat{S}(u)\right)^{\rho} (1 - \hat{S}(u))^{\gamma} \text{ for } \rho \geq 0, \gamma \geq 0$$

The standard log-rank test is a special case of the FH test with $\rho = \gamma = 0$. When $\rho = 0$ and γ > 0, the FH(0, 1) test assigns more weight to events occurring later with increasing u. Change in weight becomes smaller and smaller at the tail of survival curve. With mature follow up in a trial in the metastatic setting, the weights for events occurring approaching the database cut off tends to be stable.

The weighted log-rank statistic at a given calendar time *t* can be written as

$$W(t) = \int_0^t \hat{Q}(t, u) \frac{Y_1(u)Y_2(u)}{Y_1(u) + Y_2(u)} \left[\frac{dN_1(u)}{Y_1(u)} - \frac{dN_2(u)}{Y_2(u)}\right]$$

where $\hat{Q}(t, u)$ is the weighting function as estimated at calendar time *t* for survival time *u* based on pooled data, and calendar time *t* here means the time elapse since the first subject randomized.

4.2.3 Secondary endpoints

4.2.3.1 Progression free survival (PFS)

The secondary PFS analysis will also be based on the programmatically derived RECIST 1.1 using the Investigator tumor assessments using similar methodology as described for primary OS endpoint in Section 4.2.2. The analysis will be performed in the FAS using a stratified log-rank test, adjusting for disease status and primary tumor location. The effect of Arm A versus Arm B will be estimated by the HR together with its corresponding 95% CI from stratified Cox proportional hazards model. PFS at month 6, month 9, month 12 and month 24 will also be summarized (using the Kaplan-Meier curve) and presented by treatment arm.

Kaplan-Meier plots of PFS will be presented by treatment arm. Summaries of the number and percentage of subjects experiencing a PFS event and the type of event (RECIST 1.1 or death) will be provided along with median PFS for each treatment.

Sensitivity and supplemental analysis

The following sensitivity and supplemental analysis will be performed:

Evaluation-time bias

Sensitivity analyses will be performed to assess possible evaluation-time bias that may be introduced if scans are not performed at the protocol-scheduled timepoints. The midpoint between the time of progression and the previous evaluable RECIST assessment will be analyzed using a stratified log-rank test as described for PFS analysis above. For subjects whose death was treated as PFS event, the date of death will be used to derive the PFS time used in the analysis. This approach has been shown to be robust even in highly asymmetric assessment schedules (Sun and Chen, 2010).

Attrition bias

Attrition bias will be assessed by repeating the PFS analysis except that the actual PFS event times, rather than the censored times, of subjects who progressed or died in the absence of progression immediately following 2 or more non-evaluable tumor assessments will be included. In addition, subjects who take subsequent therapy prior to progression or death will be censored at their last evaluable assessment prior to taking the subsequent therapy. This analysis will be supported by a Kaplan-Meier plot of the time to censoring where the censoring indicator of the PFS analysis is reversed.

Effect of covariates on the HR estimate (Cox proportional hazards model)

Cox-proportional hazards modelling will be used to assess the effect of covariates on the HR estimate. A model will be constructed, containing treatment, the stratification factors and the following covariates age, sex, race, ECOG and BTC. Where ECOG status is from screening.

Interactions between treatment and stratification factors will also be tested to rule out any qualitative interaction using the approach of (Gail and Simon, 1985).

Subgroup analysis

Subgroup analyses will be conducted comparing PFS (per RECIST 1.1 using Investigator assessments) between Arm A and Arm B in the subgroups of the FAS, as specified in Section 4.2.2.

4.2.3.2 Objective response rate (ORR)

The ORR will be based on the site investigator RECIST 1.1 data and using all scans regardless of whether they were scheduled or not. Only confirmed responses will be reported for ORR at IA-2 and final analysis. The ORR will be compared between treatment A versus treatment B using a stratified Cochran-Mantel Haenszel (CMH) test as the primary analysis. The CMH test will be stratified using the same stratification factors as the primary endpoint. The results of the analysis will be presented in terms of a odds ratio and p-value. The odds ratio and p-value will be obtained using SAS PROC FREQ and the CMH test option. The STRATUM variable used in the TABLE statement will be based on primary tumor location and disease status.

As a sensitivity analysis ORR will be analyzed using logistic regression models adjusting for the same stratification factors as the primary endpoint as covariates in the model. The results of the analysis will be presented in terms of an odds ratio (an odds ratio greater than 1 will favor treatment A) together with its associated profile likelihood 95% CI (e.g. using the option 'LRCI' in SAS procedure GENMOD) and p-value (based on twice the change in log-likelihood resulting from the addition of a treatment factor to the model).

The ORR analysis will be performed in the subset of subjects in the FAS who had measurable disease at baseline. A sensitivity analysis will be produced in all subjects from FAS.

Summaries will be produced that present the number and percentage of subjects with a tumor response (CR/PR). Overall visit response data will be listed for all subjects (i.e, the FAS). For each treatment arm, best objective response (BoR) will be summarized by n (%) for each category (CR, PR, SD, PD and NE). No formal statistical analyses are planned for BoR.

Summaries of ORR/BoR will be carried out primarily for subjects with at least one visit response of CR or PR and repeated for subject with confirmed response (i.e. at least 1 visit response of CR or PR that is subsequently confirmed on a subsequent scan with CR or PR).

At IA-1 the ORR (for both confirmed and unconfirmed responses) in FAS-32w will be summarized in the following way both according to Investigator and BICR assessment per RECIST 1.1, while BICR being of primary interest: Point estimates and their exact Clopper-Pearson 95% CI of ORR will be produced by treatment. An exploratory p-value from CMH test will also be included. The analysis of confirmed ORR by BICR in FAS-32w with a measurable disease at baseline per BICR will be the primary analysis for IA-1. It will be repeated on a subset of subjects from FAS-32w. Supportive analyses of unconfirmed response rates will also be produced. ORR analyses according to Investigator will only be performed in FAS-32w. ORR Investigator analyses will be performed in subjects from FAS-32w.

ORR subgroup analysis will also be conducted for both final and interim analyses using subgroups as specified in Section 4.2.2. For IA-1 it will be using the BICR data for FAS-32w as well as for FAS-32w with a measurable disease at baseline per BICR subset, and will not include subgroups for PD-1L and MSI status.

4.2.3.3 Duration of response (DoR)

DoR will be summarized by treatment group. KM plots of DoR based on the Investigator assessment of RECIST 1.1 will be presented. Median DoR will also be summarized and calculated from the KM curve. In addition, the number of censored responders (split by lost to follow-up and still in response), number (%) of subjects with DoR ≥ 3 months, ≥ 6 months, ≥ 9 months and ≥ 12 months and time to onset of response from randomization will be summarized. Further descriptive summaries of DoR will be produced by subgroups (disease status, primary tumor location and region as described in OS section 4.2.2). Swimmer plots that clearly show the profile of each subject who responds will also be produced. All of the above outputs will be produced separately for both unconfirmed (IA-1 only) and confirmed responses apart from the swimmer plots which will include both unconfirmed and confirmed responses.

At IA-1, the DoR (for both confirmed and unconfirmed responses) will be summarized for subjects from FAS-32w, both according to Investigator and BICR assessment per RECIST 1.1 with BICR assessments of primary interest. For IA-1 summaries of confirmed responses will be considered primary analysis with unconfirmed responses supportive. Subgroup analyses will also be conducted using the BICR data for subjects from FAS-32w for subgroups as defined in Section 4.2.2.

DoR will be summarized for subjects with a measurable disease at baseline. If responders are found in subjects with non-measurable disease, the DoR may also be summarized for ITT population separately as a sensitivity. At IA-1 it will only be summarized for subjects with measurable disease at baseline.

4.2.3.4 Best objective response (BoR)

For each treatment arm, best objective response (BoR) will be summarized by n (%) for each category (CR, PR, SD, PD, NED and NE), both for confirmed and unconfirmed (for IA-1 only) response. No formal statistical analyses are planned for BoR. BoR analyses will be performed in subjects from FAS with a measurable disease at baseline.

At IA-1 the BoR BICR analyses will be performed in subjects from FAS-32w with a measurable disease at baseline per BICR, and repeated as a sensitivity analysis in a subset of FAS-32w. BoR Investigator analyses will be done in subjects from FAS-32w.

At IA-1 comparison of BoR by Investigator assessment and BoR by BICR assessment will be summarized by treatment group in a subset of subjects from FAS-32w with a measurable disease at baseline per Investigator or BICR, and it will be repeated as a sensitivity analysis in a subset of FAS-32w.

4.2.3.5 Disease control rate (DCR)

The DCR, DCR-24w, DCR-32w, and DCR-48w based on Investigator assessments per RECIST 1.1 will be summarized (i.e., number of subjects [%]) per treatment arm.

For IA-1 DCR based on BICR assessments according to RECIST 1.1 will also be produced and analysis will be performed in FAS-32w.

4.2.3.6 Percentage change in tumor size

Descriptive statistics will be provided for tumor size and percentage change from baseline in tumor size by treatment arm and visit for subjects in the FAS, both according to Investigator and BICR assessment per RECIST 1.1 with BICR assessments of primary interest. The best percentage change from baseline in tumor size will also be summarized. A waterfall plot will be included of best percentage change from baseline tumor size (sum of target lesion size) presenting each subject as a separate bar, with the bars ordered from the largest increase to the largest decrease. Reference lines at the –30% and +20% change in TL tumor size level will be added to the plots, which correspond with the definition of 'partial response' and 'progressive disease' respectively. The scale in these plots will be fixed to be from -100 to 100 to avoid presenting extreme values. All progressions will be marked with a '●' and imputed values are clearly marked with '*'.

For IA-1 summaries will also be produced in FAS-32w with BICR assessments being of primary interest.

Refer to Section 3.2.7 or the derivation of tumor size.

4.2.3.7 Additional supportive analyses

The following summary of RECIST assessments will also be provided.

The number of subjects prematurely censored will be summarized by treatment arm together with baseline prognostic factors of the prematurely censored subjects. A subject is defined as prematurely censored if they have not progressed (or died in the absence of progression) and the latest scan prior to DCO was more than one scheduled tumor assessment interval plus 2 weeks prior to the DCO date.

Additionally, summary statistics will be given for the number of days from censoring to DCO for all censored subjects.

A summary of the duration of follow-up will be presented using median time from randomization to date of censoring (date last known to have not progressed) in censored (not progressed) subjects only, presented by treatment group.

Additionally, summary statistics for the number of weeks between the time of progression and the last evaluable RECIST assessment prior to progression will be presented for each treatment group.

Summaries of the number and percentage of subjects who miss two or more consecutive RECIST assessments will be presented for each treatment group.

In addition, a summary of new lesions (i.e. sites of new lesions) will be produced by treatment arm.

4.2.4 Patient reported outcomes (PROs)

All patient reported outcomes will be summarized for the PRO analysis set.

Compliance rates summarizing questionnaire completion at each visit will be tabulated.

4.2.4.1 EORTC QLQ-C30

Time to deterioration

The primary assessment of symptoms, impacts, and global health status/QoL will focus on time to deterioration (TTD), which will be analyzed using a stratified log-rank tests adjusting for disease status and primary tumor location providing a p-value and stratified Cox proportional hazard model providing hazard ratio (HR) (95% CI) as described for the PFS endpoint. Separate analyses will be conducted for time to deterioration of global health status/QoL, function (including physical, role, cognitive, emotional and social), multi-term symptoms (including fatigue, pain and nausea/vomiting), and single items (dyspnoea, insomnia, appetite loss, constipation and diarrhoea). The effect of durvalumab therapy versus placebo will be estimated by the HR together with its corresponding CI and p-value. Kaplan-Meier plots will be presented by treatment group. Summaries of the number and percentage of subjects who have an event as well as who were censored will be provided along with the medians for each treatment.

Adjusted mean change from baseline

Additional analyses of global health status/QoL, impacts, and symptoms will focus on comparing mean change from baseline in the global health status/QoL, functions (physical, role, cognitive, social, and emotional), multi-term symptoms (fatigue, pain and nausea/vomiting), and single items (dyspnoea, insomnia, appetite loss, constipation and diarrhoea) score between treatment groups. The analysis population for mean change in global health status/QoL, impacts, or symptoms data will be the PRO analysis set with an evaluable baseline assessment and at least 1 evaluable post-baseline on treatment assessment. Any assessments taken after last dose of study treatment will be excluded from analysis.

Change from baseline will be derived using a mixed model repeated measures (MMRM) analysis of all the post-baseline scores for each visit. The model will include treatment, visit, and treatment-by-visit interaction as explanatory variables and the baseline score and the baseline score by visit interaction as covariates. Adjusted mean change from baseline estimates per treatment group and corresponding 95% CIs will be presented along with an overall estimate of the treatment difference, 95% CI, and p-value.

Response by visit and best overall response

Summary tables of visit responses for each EORTC QLQ-C30 scale/item score (global health status/QoL, 5 functions, and all symptoms [fatigue, pain, nausea/vomiting, dyspnoea, insomnia, appetite loss, constipation and diarrhoea]) and for each visit (improvement, deterioration, and no change) will be presented by treatment group. In addition, summary tables of the best overall response will be provided for the following domains by treatment group: global health status/QoL, function (physical, role, cognitive, social, and emotional), multi-term symptoms (fatigue and pain), and single items (appetite loss and insomnia). Occurrence of symptom, impacts, and QoL/function improvement based on best overall response will be compared between treatment groups using a logistic regression model as described for ORR. The odds ratio, p-value, and 95% CI will be presented graphically on a forest plot.

Change from baseline

Finally, summaries of absolute and unadjusted change from baseline values of each EORTC QLQ-C30 scale/item will be reported by visit for each treatment group. Graphical presentations may also be produced as appropriate.

4.2.4.2 EORTC QLQ-BIL21

Time to deterioration

The primary assessment of TTD, as described for the EORTC QLQ-C30, will be evaluated for single-item abdominal pain (item 42), pruritus (item 36), jaundice (item 35) and weight loss (item 51) and symptoms (eating scale, jaundice, pain, anxiety and tiredness) of the EORTC QLQ-BIL21. TTD will be presented using a Kaplan-Meier plot as well as the HR and corresponding 95% CI from stratified Cox proportional hazard model and p-value from stratified log rank test. Summaries of the number and percentage of subjects experiencing a clinically meaningful deterioration or death and the median TTD will also be provided for each treatment group.

Mean change from baseline

Additionally, comparing mean change from baseline using the MMRM as described for the EORTC QLQ-C30 will be repeated for single-item abdominal pain (item 42), pruritus (item 36), jaundice (item 35) and weight loss (item 51) and symptoms (eating scale, jaundice, pain, anxiety and tiredness) of the of the EORTC QLQ-BIL21. All assumptions and outputs as described for the EORTC QLQ-C30 are applicable.

Response by visit and best overall response

Summary tables of visit responses for single-item abdominal pain (item 42), pruritus (item 36), jaundice (item 35) and weight loss (item 51) and symptoms (eating scale, jaundice, pain, anxiety and tiredness) will be presented by treatment group. In addition, for each visit, improvement, deterioration, and no change will be presented by treatment group. In addition, summary tables of best overall response will be provided. Occurrence of improvement based on best overall response will be compared between treatment groups using a logistic regression model. The odds ratio, p-value, and 95% CI will be presented graphically on a forest plot.

Change from baseline

As described for the EORTC QLQ-C30, summaries of absolute and unadjusted change from baseline values of each EORTC QLQ-BIL21 scale/item will be reported by visit for each treatment group. Graphical presentations may also be produced as appropriate.

4.2.4.3 PRO-CTCAE

Data from the PRO-CTCAE will be summarized by treatment group. The number and percentage of subjects with each level of response for each CTCAE item at baseline and over time will be summarized. A bar chart of the incidence by visit may be presented for each CTCAE. Further summaries to explore the data (i.e., the severity of symptoms) may be produced if needed.

Similar to PRO-CTCAE analysis, item 49 of the EORTC QLQ-BIL21 ("To what extent have you been troubled with side-effects from your treatment?") assessing patient's global impression of treatment tolerability will be evaluated and graphically presented to complement exploratory findings of the PRO-CTCAE.

4.2.4.4 PGIS

Responses for PGIS will be summarized descriptively as number of subjects and corresponding percentage for each category in questionnaire at each visit by treatment group.

4.2.4.5 EQ-5D-5L

Descriptive statistics will be calculated for each scheduled visit/timepoint in the study, for each study arm, and as a total. This will report the number of subjects, the number of EQ-5D questionnaires completed at each visit, and the number and proportion responding to each dimension of the EQ-5D-5L. Additionally, summary will be reported for the EQ-5D index score and the EQ-VAS score, as well as the change from baseline for the EQ-5D index score and the EQ-VAS score.

Graphical plots of the mean EQ-5D index score and EQ-VAS score, including change from baseline, and associated 95% CI by scheduled visits/timepoints in the study may be produced. To support submissions to payers, additional analyses may be undertaken, and these will be outlined in a separate Payer Analysis Plan (PAP).

4.2.5 Healthcare resource use

The HOSPAD module is for all non-study protocol-related hospital admissions; any routine hospital visits for study protocol-related requirements do not need to be captured. This would include providing descriptive statistics as appropriate, including means, median, ranges or frequencies, and percentages.

To support submissions to payers, additional analyses may be undertaken, and these will be outlined in a separate PAP.

4.2.6 Safety data

Safety and tolerability data from all cycles of treatment will be combined and will be presented by treatment arm using the SAF. Safety summaries will be descriptive only. No formal statistical analyses will be performed on the safety variables.

The following sections describe the planned safety summaries for AEs, vital signs, laboratory parameters, ECG and WHO performance status. However, additional safety summaries (not specified in this SAP) may need to be produced to aid interpretation of the safety data.

4.2.6.1 General considerations for safety and PRO assessments

Time windows for safety data and PRO assessments

Time windows will be defined for all presentations of safety data that summarize values by visit according to the following conventions:

- Safety and PRO data study day will reference 1st dose. For subjects randomized and not treated, randomization date will be used instead to assign study day for PRO data.
- The time windows should be exhaustive so that data recorded at any time point (scheduled or unscheduled) has the potential to be summarized. Inclusion within the time window should be based on the actual data and not the intended date of the visit.

- The window for visits following baseline will be constructed in such a way that the upper limit of the interval falls halfway between the two visits (the lower limit of the first post baseline visit will be Day 2), see Appendix B for all safety data and PRO visit windows. If an even number of days exist between two consecutive visits, then the upper limit will be taken as the midpoint value minus 1 day. For summaries showing the maximum or minimum values, the maximum/minimum value recorded on treatment (as defined in Section 4.2.6.5) will be used (regardless of where it falls in an interval).

- Listings will display all values contributing to a time point for a subject.

- For visit-based summaries:

  - If there is more than one value per subject within a time window, then the closest value to the scheduled visit date will be summarized. If the values are equidistant from the nominal visit date, then the earlier value will be used. Data listings will highlight the values used in the summary table, wherever feasible. Note: In summaries of extreme values, all post-baseline values collected are used including those collected at unscheduled visits regardless of which value is closest to the scheduled visit date.

  - Visit data will only be summarized if the number of observations is ≥20 in at least one treatment arm.

- For summaries at subject level, all values will be included when deriving a subject level statistic such as a maximum regardless of whether they appear in the corresponding visit-based summary.

Handling of missing data

Missing safety data will generally not be imputed. However, safety assessments of the form of “<x” (i.e., below the lower limit of quantification) or “>x” (i.e., above the upper limit of quantification) will be imputed as “x” in the calculation of summary statistics but will be displayed as “<x” or “>x” in the listings.

For missing start dates for AEs and concomitant medications/procedures, the following will be applied:

- Missing day: Impute the 1st of the month unless month is the same as month of the first dose of study drug then impute first dose date.

- Missing day and month: Impute 1st January unless year is the same as first dose date then impute first dose date.

- Completely missing date: Impute first dose date unless the end date suggests it could have started prior to this in which case impute the 1st January of the same year as the end date.

When imputing a start date, ensure that the new imputed date is sensible e.g., prior to the end date of the AE.

For missing stop dates of AEs or concomitant medications/procedures, the following will be applied:

- Missing day: Impute the last day of the month unless month is the same as month of last dose of study drug then impute last dose date.
- Missing day and month: Impute 31st December unless year is the same as last dose date then impute last dose date.
- Completely missing: If an AE/medication has a completely missing end date then it will be treated as ongoing. Flags will be retained in the database indicating where any programmatic imputation has been applied, and in such cases, any durations would not be calculated.

If a subject is known to have died where only a partial death date is available, then the date of death will be imputed as the latest of the last date known to be alive +1 from the database and the death date using the available information provided:

- Missing day only: Using the 1st of the month.
- Missing day and month: Using the 1st January.

Subjects with a partial date of birth (i.e., for those countries where year of birth only is given) will have 1st of the month imputed if the day is missing, and 1st Jan imputed if the day and month is missing.

For partial subsequent anti-cancer therapy dates, the following will be applied:

- Missing day: If the month is the same as treatment end date then impute to the day after treatment, otherwise first day of the month.
- Missing day and month: If year is the same as treatment end date then impute to the day after treatment, otherwise 1st January of the same year as anti-cancer therapy date.

4.2.6.2 Adverse events

All AEs, both in terms of current MedDRA preferred term and CTCAE grade, will be summarized descriptively by count (n) and percentage (%) for each treatment group. Any AE occurring before randomized treatment (i.e. before the administration of the first infusion on Study Day 1) will be included in the AE listings, but will not be included in the summary tables (unless otherwise stated). These will be referred to as 'pre-treatment'. However, any AE occurring before the administration of the first dose on Study Day 1 that increases in severity after the first dose will be regarded as treatment emergent and thus will be included in the summary tables. Note: If an AE is not worse than baseline (pre-dose) severity then it will not be classified as TEAE.

AEs observed up until 90 days following last dose of the study treatment or until the initiation of the first subsequent anti-cancer therapy following discontinuation of study treatment (whichever occurs first) will be used for reporting of all the AE summary tables. This will more accurately depict AEs attributable to study treatment only as some of AEs up to 90 days following discontinuation of the study treatment are likely to be attributable to subsequent therapy.

However, to assess the longer-term toxicity profile, limited AE summaries may also be produced containing AEs observed up until 90 days following discontinuation of the durvalumab plus gemcitabine/cisplatin combination therapy or placebo plus gemcitabine/cisplatin combination therapy (i.e. without taking subsequent therapy into account). Any events in this period that occur after a subject has received further therapy for cancer (following discontinuation of study treatment) will be flagged in the data listings.
A separate listing of AEs occurring more than 90 days after discontinuation of study treatment or after initiation of subsequent cancer therapy will be produced. These events will not be included in AE summaries.

All reported AEs will be listed along with the date of onset, date of resolution (if AE is resolved) and investigator's assessment of severity and relationship to study drug. Frequencies and percentages of subjects reporting each preferred term (PT) will be presented (i.e. multiple events per subject will not be accounted for apart from on the episode level summaries which may be produced).

Summary information (the number and percent of subjects by system organ class and PT separated by treatment group) will be tabulated for:

- All AEs
- All AEs possibly related to any study medication (as determined by the reporting

- investigator)
- AEs with CTCAE grade 3 or 4
- AEs with CTCAE grade 3 or 4, possibly related to any study medication (as determined by the reporting investigator)
- Most common AEs
- Most common AEs with CTCAE grade 3 or 4
- AEs with outcome of death
- AEs with outcome of death possibly related to any study medication (as determined by the reporting investigator)
- All SAEs
- All SAEs possibly related to any study medication (as determined by the reporting investigator)
- AEs leading to discontinuation of any study medication
- AEs leading to discontinuation of any study medication, possibly related to any study medication (as determined by the reporting investigator)
- AEs leading to discontinuation of durvalumab/placebo
- AEs leading to discontinuation of gemcitabine and / or cisplatin
- AEs leading to dose interruption/reduction
- AEs leading to dose interruption/delay of durvalumab/placebo
- AEs leading to dose interruption/reduction of Gemcitabine and / or Cisplatin

An overall summary of the number and percentage of subjects in each category will be presented. For the truncated AE tables of most common AEs, all events that occur in at least 5% of subjects in one of the treatment arms will be summarized by preferred term, by decreasing frequency. This cut-off may be modified after review of the data. When applying a cut-off (e.g., 5%), the raw percentage should be compared to the cut-off, no rounding should be applied first (i.e., an AE with frequency 4.9% will not appear if the cut-off is 5%).

Each AE event rate (per 100 subject years) will also be summarized by preferred term within each system order class for the output summarizing all AEs. For each preferred term, the event rate is defined as the number of subjects with that AE divided by the total treatment duration (days) of randomized treatment summed over subjects and then multiplied by 365.25 x 100 to present in terms of per 100 subject years.

AEs will be assigned CTCAE grades and summaries of the number and percentage of subjects will be provided by maximum reported CTCAE grade, system organ class and preferred term. For each AE, time to first onset of the AE from date of first dose may be presented in the listing.

Deaths

A summary of deaths will be provided with number and percentage of subjects, categorized as below, where TEAE = AEs with onset <= 90 days from last dose and before initiation subsequent anti-cancer therapy

- Total number of deaths (regardless of date of death)
- Deaths related to disease under investigation only
- TEAE with outcome of death only
- AE outcome of death only and onset date > 90 days following last dose of study medication or initiation of subsequent anti-cancer therapy (whichever is earlier)
- Death related to disease under investigation and TEAE with outcome of death
- Death related to disease under investigation and AE with outcome of death > 90 days after last dose of study medication or ≥ date of subsequent anti-cancer therapy, whichever occurs first
- Subjects with unknown reason for death.
- Other deaths.

Other significant adverse events

The number and percent of subjects by treatment group with other significant adverse events will be summarized by Standardized MedDRA Queries (SMQ), MedDRA preferred term and maximum reported CTCAE grade. Separate outputs will be produced for Hepatic disorder SMQs Biliary disorder SMQs and Hematopoietic SMQs. The analysis will be repeated for adverse events related to any study medication.

4.2.6.3 Adverse events of special interest (AESI) and possible interest (AEPI)

Preferred terms used to identify AESI/AEPI, as defined in Section 3.5.6 will be listed before database lock (DBL) and documented in the Trial Master File. Grouped summary tables for certain MedDRA preferred terms will be produced and may also show the individual preferred terms which constitute each AESI/AEPI grouping. Groupings will be based on preferred terms provided by the medical team prior to DBL, and a listing of the preferred terms in each grouping will be provided.

Summaries of the above-mentioned grouped AE categories will include number (%) of subjects who have:

- Any AESI/AEPI
- Any AESI/AEPI by SOC, PT and maximum CTCAE grade
- Any AESI/AEPI by SOC, PT with CTCAE grade 3 or 4
- Any serious AESI/AEPI
- Any AESI/AEPI with outcome of death
- At least one AESI/AEPI possibly related to any study medication (as determined by the reporting investigator)
- Any AESI/AEPI leading to concomitant medication use (corticosteroids)
- Any AESI/AEPI leading to concomitant medication use (high dose steroids)
- Any AESI/AEPI leading to concomitant medication use (endocrine therapy)
- Any AESI/AEPI leading to concomitant medication use (other immunosuppressants)
- At least one AESI leading to discontinuation of any study medication
- At least one AESI leading to discontinuation of Durvalumab/Placebo
- At least one AESI leading to discontinuation of Gemcitabine and /or Cisplatin

An overall AESI/AEPI summary will be presented, including number and percentage of subjects in each of these categories. Any AESI/AEPI presented by outcome will also be provided.

Additionally, summaries will include time to onset of first CTCAE grade 3 or 4. Time to onset of first AE for each grouped term and preferred term within it will also be produced.

Infusion reaction adverse events

The number and percent of subjects with infusion reaction adverse events will be summarized by system organ class and PT by treatment group.

Immune-mediated Adverse events (imAEs)

The imAEs (as classified by the Sponsor) will be summarized in the similar manner as for the summaries for AESI/AEPI described above. The Sponsor will be responsible for producing these summaries.

Interstitial Lung Disease (ILD) events

The following summaries will be provided for ILD events:

- ILD events - list of preferred terms
- ILD events by AESI/AEPI grouped term, preferred term and maximum reported CTCAE grade
- ILD events with outcome of death by AESI/AEPI grouped term and preferred term
- ILD events leading to discontinuation of any study medication by AESI/AEPI grouped term and preferred term
- ILD events leading to discontinuation of Durvalumab/Placebo by AESI/AEPI grouped term and preferred term
- ILD events leading to discontinuation of Gemcitabine and / or Cisplatin by AESI/AEPI grouped term and preferred term

In addition, a listing of key information for ILD events will be provided.

4.2.6.4 Exposure

Exposure will be summarized for the SAF. The following summaries will be produced:

- Total exposure.
- Actual exposure (durvalumab or matching placebo, gemcitabine or cisplatin).

- RDI (durvalumab or matching placebo, gemcitabine or cisplatin).
- Number of cycles received (durvalumab or matching placebo, gemcitabine or cisplatin).
- Summary of duration of exposure (durvalumab or matching placebo, gemcitabine or cisplatin).
- Summary of interruptions and reductions for durvalumab or matching placebo, gemcitabine or cisplatin. Dose interruptions will be based on investigator dosing decisions.

4.2.6.5 Laboratory measurements

Laboratory data obtained until 90 days after the last dose of study treatment or until the initiation of the first subsequent anti-cancer therapy following discontinuation of study treatment (whichever occurs first) will be used for reporting. This will more accurately depict laboratory toxicities attributable to study treatment only as a number of toxicities up to 90 days following discontinuation of study treatment are likely to be attributable to subsequent anti-cancer therapy.

Data summaries and listings will be provided by AZ preferred units.

All laboratory data will be listed. Flags will be applied to values falling outside – reference ranges (which will be explicitly noted on these listings where applicable), and to values for which CTCAE grading applies.

Scatter plots (shift plots) of baseline to maximum/minimum values (as appropriate) on treatment (i.e., on treatment is defined as data collected between the start of treatment and the relevant follow-up period following the last dose of study treatment) may be produced for certain parameters if warranted after data review.

Box-plots of absolute values by week, and box-plots of change from baseline by week, may be presented for certain parameters if warranted after data review.

Shift tables of laboratory values by worst common toxicity criteria (CTCAE) grade will be produced, and for specific parameters separate shift tables indicating hyper- and hypo-directionality of change will be produced. The laboratory parameters for which CTCAE grade shift outputs will be produced are:

- Hematology: Hemoglobin, Leukocytes, Lymphocytes (absolute count), Neutrophils (absolute count), Platelets

- Clinical Chemistry: ALT, AST, Albumin, Alkaline Phosphatase (ALP), Total bilirubin, Magnesium (hypo- and hyper-), Sodium (hypo- and hyper-), Potassium (hypo- and hyper-), Corrected Calcium (hypo- and hyper-), Glucose (hypo- and hyper-), Gamma-glutamyl transferase, Creatinine.

For parameters with no CTCAE grading that are listed in the CSP, shift tables from baseline to worst value on treatment will be provided. Additional summaries will include a shift table of urinalysis (Bilirubin, Blood, Glucose, Ketones, Protein) comparing baseline value to maximum on treatment value.

The denominator used in laboratory summaries of CTCAE grades will only include evaluable subjects. If a CTCAE criterion involved a change from baseline, evaluable subjects are those who have both a pre-dose and at least 1 post-dose value recorded. If a CTCAE criterion does not consider changes from baseline. Evaluable subjects are those who have at least 1 post-dose value recorded.

A shift table with changes from baseline of CrCl calculated using Cockroft-Gault formula will be created:

- Normal: Glomerular filtration rate (GFR) >= 90 mL/min;
- Mild Impairment: GFR >= 60 - < 90 mL/min;
- Moderate Impairment: GFR >= 30 - < 60 mL/min;
- Severe Impairment: GFR >= 15 - < 30 mL/min;
- Kidney Failure: GFR < 15 mL/min.

Reversibility of creatinine clearance calculated using Cockroft-Gault formula will be summarized:

- Subjects shifting into a worse renal impairment category from baseline
- Subjects whose shift from baseline was reversible and transient (reversible and transient is defined as a subsequent CrCl value that is higher than the worst CrCl value and in a better impairment category).

Hy’s law (HL)

The following summaries will include the number (%) of subjects who have:

- Elevated ALT, AST, and Total bilirubin during the study

- ALT ≥3x – ≤5x, >5x – ≤8x, >8x – ≤10x, >10x – ≤20x, and >20x ULN during the study.

- AST ≥3x – ≤5x, >5x – ≤8x, >8x – ≤10x, >10x – ≤20x, and >20x ULN during the study.

- Total bilirubin ≥2x – ≤3x, >3x – ≤5x, >5x ULN during the study.
- ALT or AST ≥3x – ≤5x, >5x – ≤8x, >8x – ≤10x, >10x – ≤20x, >20x ULN during the study.
- ALT or AST ≥3x ULN and total bilirubin ≥2x ULN during the study (potential Hy's law): the onset date of ALT or AST elevation should be prior to or on the date of total bilirubin elevation irrespective of an increase in Alkaline Phosphatase (ALP).

Narratives may be provided in the CSR for subjects who have ALT ≥3x ULN plus total bilirubin ≥2x ULN or AST ≥3x ULN plus total bilirubin ≥ 2x ULN at any visit.

Liver biochemistry test results over time for subjects with elevated ALT (i.e. ≥3x ULN) or AST (i.e. ≥3x ULN), and elevated total bilirubin (i.e. ≥2x ULN) (at any time) will be plotted.

Individual subject data where ALT or AST plus total bilirubin are elevated at any time will be listed also.

Plots of maximum post-baseline ALT and AST vs. maximum post-baseline total bilirubin, expressed as multiples of ULN, will also be produced with reference lines at 3×ULN for ALT and AST, and 2×ULN for Total bilirubin. In each plot, total bilirubin will be in the vertical axis.

Abnormal Thyroid function

Elevated thyroid stimulating hormone (TSH) will be summarized per treatment group in terms of number (%) of subjects with elevated TSH (higher than the upper normal range), low TSH (lower than lower normal range), elevated TSH post-dose and within normal range at baseline, low TSH post-dose and within normal range at baseline. Shift tables showing baseline to maximum and baseline to minimum will be produced. Additionally free T3/ free T4 data will be summarized for each TSH category (at least one free T3/ free T4 >LLN, all other T3 free/ T4 free >= LLN, free T3/ free T4 missing).

Pregnancy tests

A listing including all pregnancy test results will be produced.

4.2.6.6 Electrocardiograms

ECG data obtained up until the safety follow-up will be included in the summary tables. Absolute values and change from baseline for ECG heart rate, PR duration, QRS duration, QT duration, and RR duration may be presented.

Overall evaluation of ECG is collected in terms of normal or abnormal, and the relevance of the abnormality is termed as “clinically significant” or “not clinically significant”. ECG evaluations will be summarized using a shift table of baseline to worst evaluation on-treatment during the study if a sufficient number of ECG assessments are recorded.

4.2.6.7 Physical examination

Individual physical examination data will not be summarized.

4.2.6.8 Vital signs

Summaries for vital signs data will include all data obtained until 90 days after the last dose of study treatment. Absolute values and change from baseline for diastolic and systolic BP, pulse, respiratory rate, temperature and weight will be summarized at each visit. The denominator in vital sign data should include only those subjects with recorded data.

Box-plots for absolute values and change from baseline by week may be presented for certain vital signs parameters if warranted after data review.

4.2.6.9 WHO/ECOG performance status

All WHO/ECOG performance status data will be summarized over time. Absolute values and change from baseline for WHO/ECOG PS will be summarized at each visit.

4.2.7 Pharmacokinetic data

PK concentration data for durvalumab will be summarized for all subjects in the PK analysis set.

Serum concentrations of durvalumab will be summarized by nominal sample time using standard summary statistics for PK concentrations (geometric mean, geometric coefficient of variation, arithmetic mean, standard deviation, minimum, maximum and n). All serum concentrations will be listed.

If the data are suitable, the relationship between PK exposure and efficacy/safety parameters may be investigated graphically or using an appropriate data modelling approach.

4.2.8 Immunogenicity analysis

Immunogenicity results of all subjects will be listed regardless of ADA positive/negative status. The number and percentage of subjects who develop detectable ADA to durvalumab within each ADA response category listed in Section 3.7 will be summarized based on the

ADA analysis set. ADA titer and nAb data will be listed for samples confirmed positive for the presence of anti-durvalumab antibodies. Details for the presentation and derivation of ADA data is provided in Section 3.7. AEs in ADA positive subjects by ADA positive category will be listed.

The effect of immunogenicity on PK, efficacy and safety will be evaluated if data allow.

4.2.9 Biomarkers

The relationship of PD-L1 expression and, if applicable, of exploratory biomarkers (e.g., microsatellite instability (MSI)/mismatch repair proficiency) to clinical outcomes including but not restricted to OS, PFS, ORR and DoR will be presented. PD-L1 expression and MSI/mismatch repair proficiency will be reported in the CSR. Summaries and analyses for other exploratory biomarkers will be documented in a separate analysis plan and will be reported outside the CSR in a separate report. Baseline PD-L1 and MSI data will be listed.

PD-L1 expression (low vs. high)

Patients will provide a tumor tissue sample at screening. Tumor evaluations of PD-L1 expression are intended to be performed for all randomized patients.

PD-L1 expression will be determined by the analytically validated VENTANA PD-L1 (SP263) Assay using the TIP score method. The TIP score will be defined as the total percentage of the tumor area covered by tumor cells with PD-L1 membrane staining at any intensity and tumor-associated immune cells with any pattern of PD-L1 staining at any intensity. PD-L1-High will be defined as PD-L1 staining of any intensity in tumor cell membranes and tumor-associated immune cells covering ≥1% of tumor area. PD-L1-Low will be defined as PD-L1 staining of any intensity in tumor cell membranes and/or tumor-associated immune cells covering <1% of tumor area.

- TIP ≥1% PD-L1 is considered high expression (PD-L1-High)
- TIP <1% PD-L1 is considered low expression (PD-L1-Low)

An on-treatment specimen will also be collected to assess the pathological stage and PD-L1 status for exploratory tumor biomarker analysis.

MSI status (MSI-H vs. MSS)

Microsatellite Instability (MSI) is a genomic signature of deficient mismatch repair (dMMR) which may be associated with response to immunotherapy. MSI analysis is intended for all randomized subjects with sufficient tumor tissue remaining after PD-L1 IHC. MSI assessment will not be conducted for subjects in China.

Screening tumor tissues will be evaluated retrospectively using the FoundationOne (F1) laboratory developed test (LDT) assay. To determine MSI status, 114 microsatellite loci will be sequenced, analyzed for length variability and compiled into an overall MSI score via principal component analysis. Each subject sample will be assigned a qualitative status of MSI-High (MSI-H) or MSI-Stable (MSS) based on the guidance in Table 21 Samples with low sequence coverage (< 250X median) will be assigned a status of MSI-unknown. If there are too few subjects in MSI high group (n<5 in both treatment groups), MSI subgroup analysis will not be performed.

Table 21: MSI score range

MSI Status	**MSI Score Range**
MSI-High	≤ -8.5
MSI-Stable	> -4.0
MSI Unknown	Low sequence coverage

4.2.10 Demographic, initial diagnosis and screening or baseline characteristics data

The following will be summarized for all subjects in the FAS (unless otherwise specified) by treatment group:

- Subject disposition (including screening failures and reason for screening failure)
- Important protocol deviations
- Inclusion in analysis sets
- Demographics (age, age group <65, ≥65 - <75 and ≥75 years], sex, race and ethnicity)
- Subject characteristics at baseline (height, weight, weight group, PD-L1 expression and MSI status)
- Subject recruitment by region, country and center
- Stratification factors recorded at randomization on the IVRS and eCRF
- Previous disease-related treatment modalities

- Previous chemotherapy prior to this study
- Disease characteristics at initial diagnosis or screening (ECOG performance status, primary tumor location, histology type, tumor grade and overall disease classification)
- Virology status at baseline (No viral hepatitis, any viral hepatitis B, active viral hepatitis B (this a subset of any viral hepatitis B) and prior hepatitis C).
- Extent of disease at baseline (locally advanced and metastatic)
- TNM classification at baseline (summarized separately by initially unresectable and recurrent)
- Medical history (past and current)
- Surgical history
- Surgical history related to Biliary tract cancer
- Disallowed concomitant medications
- Allowed concomitant medications
- Post-discontinuation cancer therapy

The medications will be coded following AZ standard drug dictionary/WHO Drug dictionary as applicable.

4.2.11 Concomitant and other treatments

All concomitant and other treatment data will be listed for all subjects in the FAS.

Allowed and disallowed concomitant medications will be presented by treatment arm, ATC classification and generic term for the FAS. Subjects with the same concomitant medication/procedure multiple times will be counted once per medication/procedure. A medication/procedure that can be classified into more than once chemical and/or therapeutic subgroup will be presented in each subgroup.

Concomitant surgical procedures related to biliary tract

The number and percent of subjects by treatment group with concomitant surgical procedures related to biliary tract will be summarized by MedDRA preferred term. A list of stent or related PTs will be provided by medical experts prior to the database lock.

4.2.12 COVID-19

Depending on the extent of any impact, summaries of data relating to subjects diagnosed with COVID-19, and impact of COVID-19 on study conduct (in particular missed visits, delayed or discontinued IP, and other protocol deviations) may be generated including

- Disposition (discontinued IP due to COVID-19 and withdrew study due to COVID-19)
- Deviations (overall deviations plus if due to COVID-19 and not due to COVID-19)
- Summary of COVID-19 disruption (visit impact, drug impacted)
- Listing for subjects affected by the COVID-19 pandemic
- Listing for subjects with reported issues in the Clinical Trial Management System due to the COVID-19 pandemic

A sensitivity analysis of OS may be conducted to assess for the potential impact of COVID-19 deaths on OS. This will be assessed by repeating the OS analysis except that any subject who had a death with primary/secondary cause as COVID-19 infection will be censored at their COVID-19 infection death date. COVID-19 deaths will be identified by primary/secondary cause of death.

5 INTERIM ANALYSIS

5.1 Interim analyses

Interim safety monitoring will be conducted by an IDMC. Interim analyses will be performed for efficacy as described in the sections below.

5.1.1 ORR/DoR interim analysis (IA-1)

The first interim analysis will be performed after approximately 200 randomized subjects have had opportunity to be followed up for at least 32 weeks or the last subject has been randomized to the global cohort whichever comes later (i.e. randomized ≥32 weeks prior to IA-1 DCO). The objective is to evaluate the efficacy of durvalumab + gemcitabine and cisplatin in terms of clinical activity as measured by ORR and DoR. The analysis set will include all randomized subjects who have had the opportunity to be followed up for at least 32 weeks at the time of the IA-1 DCO (FAS-32w, i.e. randomized ≥32 weeks prior to IA-1 DCO). Although no formal comparison between arms will be performed, a nominal significance level of 0.001 will be allocated to IA-1.

At IA-1, no formal statistical testing will be performed, however an exploratory p-value from a stratified CMH test for ORR will be produced following FDA feedback. The primary endpoint for IA-1 is the confirmed ORR based on BICR as per RECIST 1.1 calculated FAS-32w with a measurable disease at baseline per BICR. It will be repeated in a subset of FAS-32w subjects. Descriptive summaries of ORR including a 2-sided exact Clopper-Pearson 95% confidence interval will be presented by treatment group. Further descriptive summaries of ORR will be produced by subgroups based on confirmed BICR assessments for the subgroups listed for OS endpoint in Section 4.2.2 except for MSI and PD-L1 status. The full list of efficacy summaries that will be produced for IA-1 is listed below. Full details of the definitions and summaries are provided in sections 3 and 4 of SAP.

- Confirmed ORR based on BICR assessment
- Confirmed ORR based on Investigator assessment
- Confirmed ORR subgroups based on BICR assessment (all subgroups except MSI and PD-L1)
- Confirmed BoR based on Investigator assessment
- Confirmed BoR based on BICR assessment
- Confirmed DoR based on Investigator assessment
- Confirmed DoR based on BICR assessment
- Confirmed DoR subgroups based on BICR assessment (disease status, primary tumor location and region)
- Confirmed DoR based on Investigator assessment KM plot
- Confirmed DoR based on BICR assessment KM plot
- Comparison of confirmed BoR for BICR vs Investigator assessment
- Percentage change from baseline in target lesion size based on BICR assessment
- Waterfall plot for best percentage change in target lesion size based on BICR assessment

If early registration is recommended the following additional summaries will also be produced for submission:

- Unconfirmed ORR based on Investigator assessments
- Unconfirmed ORR based on BICR assessments
- Unconfirmed BoR based on Investigator assessments
- Unconfirmed BoR based on BICR assessments

- Unconfirmed DoR based on Investigator assessments
- Unconfirmed DoR based on BICR assessments
- Unconfirmed DoR swimmer plot based on Investigator assessments
- Unconfirmed DoR swimmer plot based on BICR assessments
- Unconfirmed DoR KM plot based on Investigator assessments
- Unconfirmed DoR KM plot based on BICR assessments
- DCR based on Investigator assessments
- DCR based on BICR assessments
- Percentage change from baseline in target lesion size based on Investigator assessment
- Best percentage change from baseline in target lesion size based on BICR assessment
- Best percentage change from baseline in target lesion size based on Investigator assessment
- Summary of new lesions

At IA-1 ORR, BoR and DoR analyses based on Investigator assessment. ORR and BoR will be performed in FAS-32w and DoR in FAS-32w subset of subjects with measurable disease at baseline. ORR, BoR and DoR analyses based on BICR assessment will be performed in FAS-32w subset of subjects with measurable disease at baseline per BICR, and repeated for ORR and BoR in the FAS-32w as a sensitivity analysis.

The minimum efficacy criteria for IA-1 is:

The lower bound of the two-sided exact 95% CI for ORR in the durvalumab + gemcitabine and cisplatin arm is higher than the ORR point estimate of the placebo + gemcitabine and cisplatin arm.

The Unblind Review Committee (URC) will determine submission recommendation based on the following:

1. ORR - Satisfied as part of the IDMC criteria.
2. DoR – ORR improvement is greater than 15% and DoR is at least as good as control arm

   OR

   ORR improvement is < 15%, and either mDoR is at least 2.5month improvement (if available) or DoR landmark at 6 months is at least 15% greater than control arm
3. Safety – Durvalumab + gemcitabine and cisplatin safety profile in line with study drug regimen and disease state

The efficacy criteria for IA-1 will be applied to the analysis of confirmed ORR/ DoR per BICR in subset of subjects from FAS-32w with a measurable disease at baseline per BICR.

5.1.2 OS interim analysis (IA-2)

An IA for OS will be performed for superiority. The OS IA will occur when approximately 80% of the final number of OS events is expected to be reached (approximately 397 of 496 OS events). The alpha level allocated to OS will be 0.049 (two-sided). It will be controlled at the interim and final analysis accounting for the correlation structure between the test statistics at IA-2 and FA. The significance level for the OS analysis using the log-rank test at IA-2 will be calculated by specifying the information fraction using the Lan DeMets alpha spending function approximating an O'Brien-Fleming approach. The information fraction is calculated as the number of OS events at the interim analysis time-point divided by the total number of events at the final analysis time-point. For example, if 80% of OS events required at the time of the primary OS analysis are available at IA-2 (i.e.,397/496 events have occurred) and overall alpha level is 4.9%, the 2-sided significance level to be applied for the primary OS analysis at IA (IA-2) would be 2.38%. The significance level for the primary confirmatory OS analysis at FA will be determined based on the actual alpha spending at IA-2 and the correlation structure between IA-2 log-rank statistic and FA FH(0,1) statistic.

As a secondary evaluation, PFS will also be analyzed at the time of OS IA (IA-2) and FA, only if superiority is confirmed with OS at that DCO. Alpha level for PFS will be controlled at the interim and final analysis by using the Lan DeMets spending function that approximate a Pocock approach. The significance levels for the PFS analyses using the log-rank test will be calculated by specifying information fraction for each analysis. The information fraction is calculated as the number of PFS events at the analysis time-point divided by the total number of events at the final analysis time-point. For example, if 86% of PFS events expected at the time of the primary OS analysis (FA) are available at the time of IA-2 (i.e., 506/590 events have occurred), the 2-sided significance level to be applied for PFS at IA-2 would be 4.44%, and the 2-sided significance level to be applied for PFS analysis at FA would be 2.36% for the log-rank test.

5.2 Independent data monitoring committee

This study will use an external independent data monitoring committee (IDMC) to assess ongoing safety analyses as well as the interim efficacy analyses. The committee will meet approximately 6 months after the study has started to review the safety data from the study. The IDMC will meet at least every 6 months thereafter. For the interim analyses (both IA-1 and IA-2), the IDMC will review unblinded interim efficacy data as outlined above. Following each meeting, the IDMC will report to the sponsor and may recommend changes in the conduct of the study.

This committee will be composed of therapeutic area experts and biostatisticians, who are not employed by AstraZeneca and are free from conflict of interest.

Following the reviews, the IDMC will recommend whether the study should continue unchanged, be stopped, or be modified in any way. Once the IDMC has reached a recommendation, a report will be provided to AstraZeneca. The report will include the recommendation and any potential protocol amendments and will not contain any unblinding information.

The final decision to modify or stop the study will sit with the sponsor. The sponsor or IDMC may call additional meetings if at any time there is concern about the safety of the study.

Full details of the IDMC procedures and processes can be found in the IDMC Charter. The safety of all AstraZeneca/MedImmune clinical studies is closely monitored on an ongoing basis by AstraZeneca/MedImmune representatives in consultation with the Subject Safety Department. Issues identified will be addressed; this could involve, for instance, amendments to the Clinical Study Protocol and letters to investigators.

6 CHANGES OF ANALYSIS FROM PROTOCOL

All PRO analyses will be performed on PRO analysis set and not FAS as stated in CSP.

No subgroup analysis will be performed for DCR as considered not to be informative.

No sensitivity analysis will be carried out for ORR and DoR for subjects with 32 weeks follow-up as assessed by BICR, as BICR data will only be collected up until IA-1.

7 REFERENCES

Barrueto et al 2020
Barrueto L, Caminero F, Cash L et al. Resistance to Checkpoint Inhibition in Cancer Immunotherapy, Transl Oncol. 2020 Mar;13(3):100738.

Breslow, 1974
Breslow, NE. Covariance Analysis of Censored Survival Data. Biometrics 1974; 30:89-99.

Cox, 1972
Cox D R. Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological), 1972; 34(2):187-220.

Fayers et al. 2001
Fayers P, Aaronson NK, Bjordal K, Curran D, Groenvold M, on behalf of the EORTC Quality of Life Study Group. EORTC QLQ-C30 Scoring Manual: 3rd Edition 2001. Available on request.

FDA 2011
FDA Guidance for Industry: Clinical considerations for therapeutic cancer vaccines, 2011 https://www.fda.gov/media/82312/download.

Fleming and Harrington 1991
Fleming TR, Harrington DP. Counting Processes and Survival Analysis. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley and Sons Inc. 1991;New York.

Gail and Simon, 1985
Gail M, Simon R. Testing for qualitative interactions between treatment effects and subject subsets. Biometrics. 1985; 41:361-72.

Glimm et al. 2010
Glimm E, Maurer W, Bretz F. Hierarchical testing of multiple endpoints in group-sequential trials. Stat Med 2010;29(2):219-28.

He et al. 2021
He P, Koch G, Kurland J. (2021) Robust Group Sequential Design Using Weighted Logrank Tests and Practical Considerations in Immuno-oncology Trials. Manuscript submitted (under review)

Hoos 2012
Hoos A. Evolution of end points for cancer immunotherapy trials, Ann Oncol. 2012.

Karrison 2016
Karrison TG. Versatile tests for comparing survival curves based on weighted log-rank statistics. Stata Journal 2016;16.3:678-90.

Liu et al 2018
Liu S, Chu C, Rong A. Weighted log-rank test for time-to-event data in immunotherapy trials with random delayed treatment effect and cure rate. Pharm Stat. 2018;17(5):541-554. DOI: 10.1002/pst.1878.

Lin et al. 2020

Lin, Ray S., Ji Lin, Satrajit Roychoudhury, Keaven M. Anderson, Tianle Hu, Bo Huang, Larry F. Leon et al. "Alternative Analysis Methods for Time to Event Endpoints Under Nonproportional Hazards: A Comparative Analysis." Statistics in Biopharmaceutical Research 12, no. 2 (2020): 187-198.

Lan and DeMets 1983

Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika 1983; 70:659-63.

Osoba et al. 1998

Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol 1998;16(1):139-44.

Prior 2020

Prior TJ. Group sequential monitoring based on the maximum of weighted log-rank statistics with the Fleming-Harrington class of weights in oncology clinical trials. Stat Methods Med Res. DOI: 10.1177/0962280220931560.

Schoenfeld 1981

Schoenfeld D. The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika, 68 (Dec 1981), pp. 219-316.

Sun and Chen, 2010

Sun X, Chen C. Comparison of Finkelstein's Method with the Conventional Approach for Interval-Censored Data Analysis. Stat Biopharm Res. 2010; 2:97-108

The Review Committee for Guidance Development 2018

Guidance for developing cancer immunotherapy in late phase clinical trials 2018 https://www.pmda.go.jp/files/000221609.pdf.

Tsiatis 1982

Tsiatis AA. Repeated significance testing for a general class of statistics used in censored survival analysis. J Am Stat Assoc1982; 380: 855–861.

Van Hout, 2012

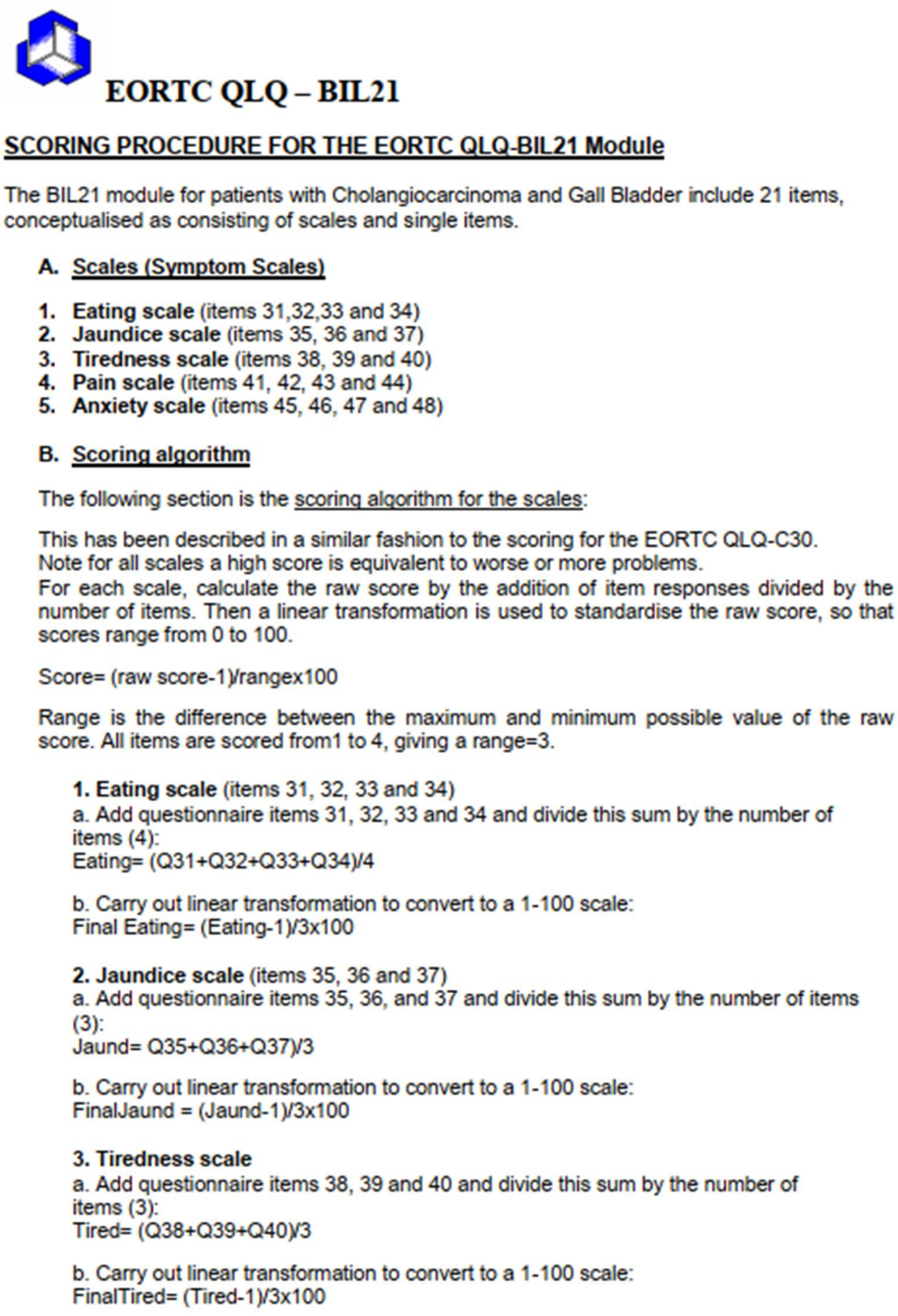
Van Hout B, Janssen MF, et al. Interim scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Health. 2012; 15(5):708-15.

Xu et al 2018
Xu Z, Park Y, Zhen B, Zhu B. Designing cancer immunotherapy trials with random treatment time-lag effect, Statistics in Medicine. 2018;37:4589–4609.

Yamaguchi et al 2014
Yamaguchi Y, Yamaue H, Okusaka T et al. Guidance for peptide vaccines for the treatment of cancer. The Committee of Guidance for Peptide Vaccines for the Treatment of Cancer, The Japanese Society for Biological Therapy. Cancer Sci 105 (2014) 924– 931.

8 APPENDICES

Appendix A EORTC QLQ – BIL 21 Scoring Procedure



EORTC QLQ – BIL21

SCORING PROCEDURE FOR THE EORTC QLQ-BIL21 Module

The BIL21 module for patients with Cholangiocarcinoma and Gall Bladder include 21 items, conceptualised as consisting of scales and single items.

A. Scales (Symptom Scales)

1. **Eating scale** (items 31,32,33 and 34)
2. **Jaundice scale** (items 35, 36 and 37)
3. **Tiredness scale** (items 38, 39 and 40)
4. **Pain scale** (items 41, 42, 43 and 44)
5. **Anxiety scale** (items 45, 46, 47 and 48)

B. Scoring algorithm

The following section is the scoring algorithm for the scales:

This has been described in a similar fashion to the scoring for the EORTC QLQ-C30.
Note for all scales a high score is equivalent to worse or more problems.
For each scale, calculate the raw score by the addition of item responses divided by the number of items. Then a linear transformation is used to standardise the raw score, so that scores range from 0 to 100.

Score= (raw score-1)/rangex100

Range is the difference between the maximum and minimum possible value of the raw score. All items are scored from1 to 4, giving a range=3.

1. Eating scale (items 31, 32, 33 and 34)
a. Add questionnaire items 31, 32, 33 and 34 and divide this sum by the number of items (4):
Eating= (Q31+Q32+Q33+Q34)/4

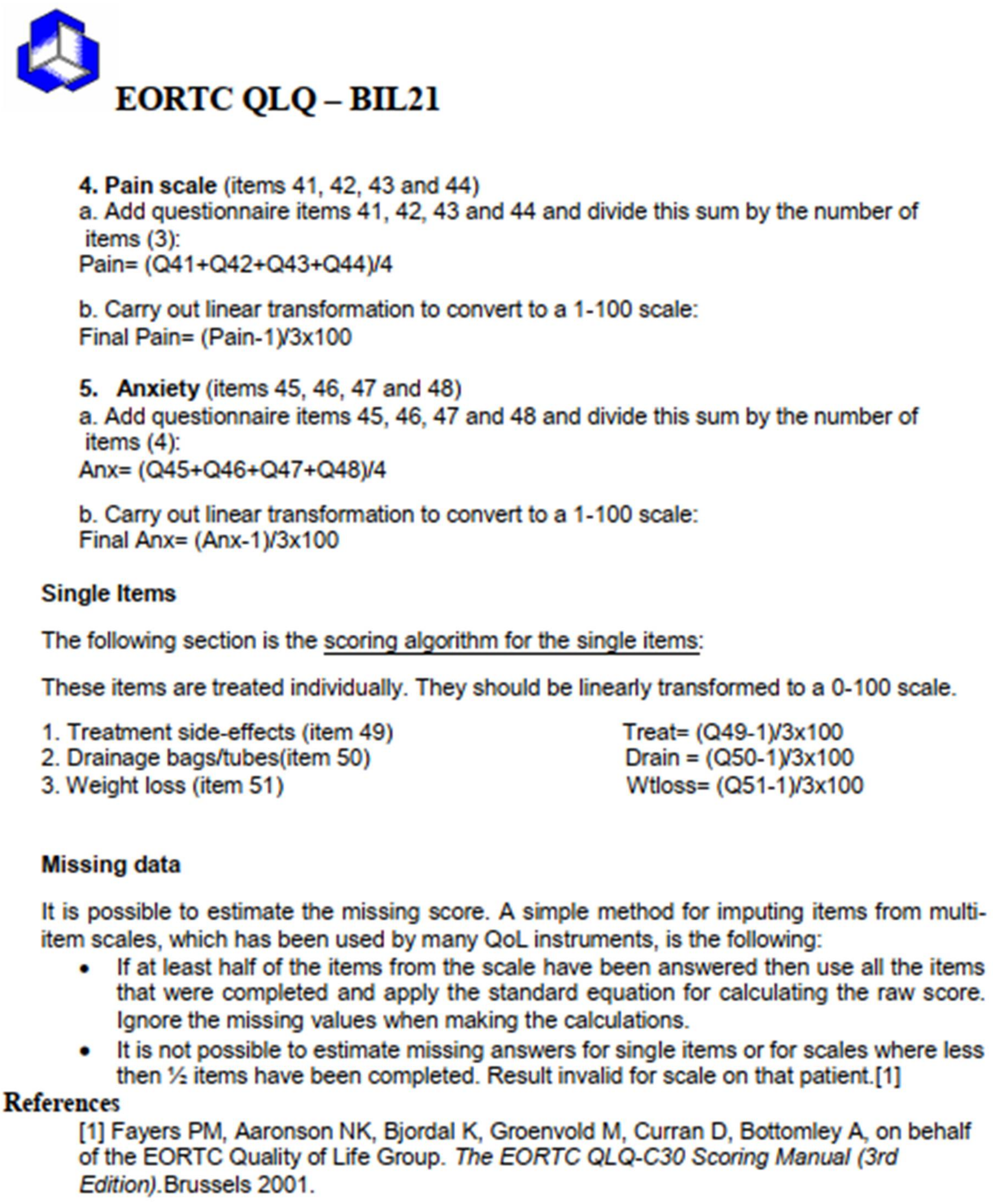
b. Carry out linear transformation to convert to a 1-100 scale:
Final Eating= (Eating-1)/3x100

2. Jaundice scale (items 35, 36 and 37)
a. Add questionnaire items 35, 36, and 37 and divide this sum by the number of items (3):
Jaund= Q35+Q36+Q37)/3

b. Carry out linear transformation to convert to a 1-100 scale:
FinalJaund = (Jaund-1)/3x100

3. Tiredness scale
a. Add questionnaire items 38, 39 and 40 and divide this sum by the number of items (3):
Tired= (Q38+Q39+Q40)/3

b. Carry out linear transformation to convert to a 1-100 scale:
FinalTired= (Tired-1)/3x100



EORTC QLQ – BIL21

4. Pain scale (items 41, 42, 43 and 44)
a. Add questionnaire items 41, 42, 43 and 44 and divide this sum by the number of items (3):
Pain= (Q41+Q42+Q43+Q44)/4

b. Carry out linear transformation to convert to a 1-100 scale:
Final Pain= (Pain-1)/3x100

5. Anxiety (items 45, 46, 47 and 48)
a. Add questionnaire items 45, 46, 47 and 48 and divide this sum by the number of items (4):
Anx= (Q45+Q46+Q47+Q48)/4

b. Carry out linear transformation to convert to a 1-100 scale:
Final Anx= (Anx-1)/3x100

Single Items

The following section is the scoring algorithm for the single items:

These items are treated individually. They should be linearly transformed to a 0-100 scale.

1. Treatment side-effects (item 49)	Treat= (Q49-1)/3x100
2. Drainage bags/tubes(item 50)	Drain = (Q50-1)/3x100
3. Weight loss (item 51)	Wtloss= (Q51-1)/3x100

Missing data

It is possible to estimate the missing score. A simple method for imputing items from multi-item scales, which has been used by many QoL instruments, is the following:

- If at least half of the items from the scale have been answered then use all the items that were completed and apply the standard equation for calculating the raw score. Ignore the missing values when making the calculations.
- It is not possible to estimate missing answers for single items or for scales where less then ½ items have been completed. Result invalid for scale on that patient.[1]

References

[1] Fayers PM, Aaronson NK, Bjordal K, Groenvold M, Curran D, Bottomley A, on behalf of the EORTC Quality of Life Group. *The EORTC QLQ-C30 Scoring Manual (3rd Edition).*Brussels 2001.

Appendix B Definition of visit windows for analysis

Table 22: Visit windows for PRO Questionnaires

PRO assessments such as, EORTC QLQ-C30, EORTC QLQ-BIL21, PGIS, EQ-5D-5L and PRO-CTCAE will use the following visit window. After Cycle 16 Day 1 administer PRO questionnaires every other cycle.

Window period	Minimum Day	Target Day	Maximum Day
Baseline	Low	1	1
Cycle 02 Day 01	2	22	32
Cycle 03 Day 01	33	43	53
Cycle 04 Day 01	54	64	74
Cycle 05 Day 01	75	85	95
Cycle 06 Day 01	96	106	116
Cycle 07 Day 01	117	127	137
Cycle 08 Day 01	138	148	158
Cycle 09 Day 01	159	169	183
Cycle 10 Day 01	184	197	211
Cycle 11 Day 01	212	225	239
Cycle 12 Day 01	240	253	267
(….and continued every cycle (4 weeks) until PD or treatment discontinuation)			
Follow-up Day 30	Last dose date + 1	Last dose date + 30	Last dose date + 45
Follow-up Month 2	Last dose date + 46	Last dose date + 60	Last dose date + 75

Follow-up Month 3	Last dose date + 76	Last dose date + 90	Last dose date + 105
Follow-up Month 4	Last dose date + 106	Last dose date + 120	Last dose date + 150
Follow-up Month 6	Last dose date + 151	Last dose date + 180	Last dose date + 210
Follow-up Month 8	Last dose date + 211	Last dose date + 240	Last dose date + 270
Follow-up Month 10	Last dose date + 271	Last dose date + 300	Last dose date + 330
Follow-up Month 12	Last dose date + 331	Last dose date + 360	Last dose date + 450
Follow-up Month 18	Last dose date + 451	Last dose date + 540	Last dose date + 630
(..and continued every 6 months)			

Table 23: Visi windows for WHO/ECOG performance status

The WHO/ECOG performance status question will follow the visit window in table 20.

Window period	**Minimum Day**	**Target Day**	**Maximum Day**
Baseline	Low	1	1
Cycle 01 Day 08	2	8	15
Cycle 02 Day 01	16	22	25
Cycle 02 Day 08	26	29	36
Cycle 03 Day 01	37	43	46
Cycle 03 Day 08	47	50	57
Cycle 04 Day 01	58	64	67
Cycle 04 Day 08	68	71	78
Cycle 05 Day 01	79	85	88

Cycle 05 Day 08	89	92	99
Cycle 06 Day 01	100	106	109
Cycle 06 Day 08	110	113	120
Cycle 07 Day 01	121	127	130
Cycle 07 Day 08	131	134	141
Cycle 08 Day 01	142	148	151
Cycle 08 Day 08	152	155	162
Cycle 09 Day 01	163	169	183
Cycle 10 Day 01	184	197	211
Cycle 11 Day 01	212	225	239
Cycle 12 Day 01	240	253	267
(…and continued every cycle (4 weeks) until treatment discontinuation)			
Follow-up Day 30	Last dose date + 1	Last dose date + 30	Last dose date + 45
Follow-up Month 2	Last dose date + 46	Last dose date + 60	Last dose date + 75
Follow-up Month 3	Last dose date + 76	Last dose date + 90	Last dose date + 105
Follow-up Month 4	Last dose date + 106	Last dose date + 120	Last dose date + 150
(…and continued every 2 months)			

Table 24: Visit windows for Laboratory measurements: Hematology and clinical chemistry

Laboratory measurements such as Hematology and clinical chemistry will use the following visit window.

Window period	Minimum Day	Target Day	Maximum Day
Baseline	Low	1	1
Cycle 01 Day 08	2	8	15
Cycle 02 Day 01	16	22	25
Cycle 02 Day 08	26	29	36
Cycle 03 Day 01	37	43	46
Cycle 03 Day 08	47	50	57
Cycle 04 Day 01	58	64	67
Cycle 04 Day 08	68	71	78
Cycle 05 Day 01	79	85	88
Cycle 05 Day 08	89	92	99
Cycle 06 Day 01	100	106	109
Cycle 06 Day 08	110	113	120
Cycle 07 Day 01	121	127	137
Cycle 08 Day 01	138	148	158
Cycle 09 Day 01	159	169	183
Cycle 10 Day 01	184	197	211
Cycle 11 Day 01	212	225	239
Cycle 12 Day 01	240	253	267

(--and continued every cycle (4 weeks) until treatment discontinuation)			
Follow-up Day 30	Last dose date + 1	Last dose date + 30	Last dose date + 45
Follow-up Month 2	Last dose date + 46	Last dose date + 60	Last dose date + 75
Follow-up Month 3	Last dose date + 76	Last dose date + 90	Last dose date + 90

Table 25: Visit windows for Laboratory measurements: Urinalysis, pregnancy, and Coagulation

Laboratory measurements such as Urinalysis and Pregnancy tests will follow the visit window as indicated in table 22. Coagulation has only screening and baseline as scheduled visits and then after that as clinically indicated.

Window period	**Minimum Day**	**Target Day**	**Maximum Day**
Baseline	Low	1	1
Cycle 02 Day 01	2	22	32
Cycle 03 Day 01	33	43	53
Cycle 04 Day 01	54	64	74
Cycle 05 Day 01	75	85	95
Cycle 06 Day 01	96	106	116
Cycle 07 Day 01	117	127	137
Cycle 08 Day 01	138	148	158
Cycle 09 Day 01	159	169	183
Cycle 10 Day 01	184	197	211

Cycle 11 Day 01	212	225	239
Cycle 12 Day 01	240	253	267
(…and continued every cycle (4 weeks) until treatment discontinuation)			
Follow-up Day 30	Last dose date + 1	Last dose date + 30	Last dose date + 45
Follow-up Month 2	Last dose date + 46	Last dose date + 60	Last dose date + 75
Follow-up Month 3	Last dose date + 76	Last dose date + 90	Last dose date + 90

Table 26: Visit windows for Vital signs and ECG

Vital signs and ECG will use the following visit window.

Window period	**Minimum Day**	**Target Day**	**Maximum Day**
Baseline	Low	1	1
Cycle 01 Day 08	2	8	15
Cycle 02 Day 01	16	22	25
Cycle 02 Day 08	26	29	36
Cycle 03 Day 01	37	43	46
Cycle 03 Day 08	47	50	57
Cycle 04 Day 01	58	64	67
Cycle 04 Day 08	68	71	78
Cycle 05 Day 01	79	85	88

Cycle 05 Day 08	89	92	99
Cycle 06 Day 01	100	106	109
Cycle 06 Day 08	110	113	120
Cycle 07 Day 01	121	127	130
Cycle 07 Day 08	131	134	141
Cycle 08 Day 01	142	148	151
Cycle 08 Day 08	152	155	162
Cycle 09 Day 01	163	169	183
Cycle 10 Day 01	184	197	211
Cycle 11 Day 01	212	225	239
Cycle 12 Day 01	240	253	267
(…and continued every cycle (4 weeks) until treatment discontinuation)			
Follow-up Day 30	Last dose date + 1	Last dose date + 30	Last dose date + 45
Follow-up Month 2	Last dose date + 46	Last dose date + 60	Last dose date + 75
Follow-up Month 3	Last dose date + 76	Last dose date + 90	Last dose date + 90

Table 27: Visit windows for RECIST data
RECIST evaluations will take place as shown in table 24.

Window period	**Minimum Day**	**Target Day**	**Maximum Day**
Baseline	Low	1	1

Cycle 03 Day 01	2	43	64
Cycle 05 Day 01	65	85	106
Cycle 07 Day 01	107	127	148
Cycle 09 Day 01	149	169	197
Cycle 11 Day 01	198	225	253
Cycle 13 Day 01	254	281	309
(…and continued every 2 cycles (8 weeks) until treatment discontinuation)			

Table 28: Visit windows for ADA data

ADA data assessments will use the following visit window.

Window period	**Minimum Day**	**Target Day**	**Maximum Day**
Baseline	Low	1	1
Cycle 05 Day 01	2	85	106
Cycle 07 Day 01	107	127	148
Follow-up Day 30	Last dose date + 1	Last dose date + 30	Last dose date + 60
Follow-up Month 3	Last dose date + 61	Last dose date + 90	Last dose date + 97

Table 29: Visit windows for PK data

PK assessments will use the following visit window.

Window period	**Minimum Day**	**Target Day**	**Maximum Day**
Baseline	Low	1	1
Cycle 02 Day 01	2	22	54
Cycle 05 Day 01	55	85	106
Cycle 07 Day 01	107	127	147
Follow-up Day 30	Last dose date + 1	Last dose date + 30	Last dose date + 60
Follow-up Month 3	Last dose date + 61	Last dose date + 90	Last dose date + 97